JOURNAL OF THEORETICAL BIOLOGY

ACADEMIC PRESS

VOLUME 1, NUMBER 3 JULY 1961

London and New York

Journal of Theoretical Biology

Volume 1, 1961: 12*l*s. 6*d*. ($17.00) Private subscriptions only: 86*s*. ($12.00)
Subscription orders should be sent to Academic Press Inc. (London) Ltd., 17 Old Queen Street, London, S.W.1, except those from Canada, Central and South America, and the U.S.A. which should be sent to Academic Press Inc., 111 Fifth Avenue, New York 3, N.Y.

J. Theoret. Biol. (1961) **1**, 289–294

On the Relation between Ionization and Biological Activity in some Sympathomimetic Amines†

HERBERT SHAPIRO

Department of Pharmacology, Woman's Medical College, Philadelphia, Pa., U.S.A.

(*Received 2 January* 1961)

By plotting values of pK_{a_1} and pK_{a_2}, the dissociation constants of the phenolic (pK_{a_1}) and amino (pK_{a_2}) group of the sympathomimetic amines, against the logarithm of the biological activity, a functional relationship having a broad maximum suggests itself. Additional active compounds at intermediate pK_a ranges, chemically related to the sympathomimetic amines, are needed to substantiate this. Similar plots for the oxidative action of the enzyme amine oxidase on these compounds yield different types of functions, one apparently linear for pK_{a_2}, and the other approximating a broadly concave downward curve for pK_{a_1}. Different sites of chemical action presumably are involved for oxidation, as contrasted to pharmacological activity. It is not to be inferred from the foregoing that the role of chemical structure is minimized. Its intent is to reassert the importance of ionization, as a molecular characteristic, in elaborating biological effects, both facets requiring examination in the elucidation of the mechanism of action of any given series of compounds. Analogously, the strength of organic acids and bases, as indicated by their dissociation constants, is altered by substituents in the organic molecule. The scatter of the points is interpreted to indicate the degree to which structure overrides ionization in the manifestation of biological effect, as well as some degree of experimental indeterminacy. The net effect of a compound will be determined by the interplay between structure and physico-chemical characteristics (Albert, 1960).

1. Introduction

The role of ionization as a determinant of the penetrability of both organic and inorganic solutes into living cells has been generally recognized and accepted for several decades. It has been enunciated in divers texts (Davson, 1951; Davson & Danielli, 1952; Höber *et al.*, 1945; Scarth & Lloyd, 1930; West, 1956). In plants, Irwin (1926), for example, demonstrated that the rate of penetration of the dye Brilliant Cresyl Blue into living *Nitella* cells was highest at the greater external pH values, indicating that the dye entered only as the undissociated molecule.

Although the above considerations have been directed toward ionization

† Aided by Grant B-1657 of the National Institutes of Health, Bethesda, Maryland.

and permeation of compounds through the cell membrane, the problem is related to actions at the cell surface, where physiological and pharmacological effects reflect the mutual interactions of the molecular structures of the biological surface or interface with the chemical agent. It is relevant here to quote Pauling (1959) ". . . the conclusion I have reached is that, almost without exception, the physiological activity of simple substances —drugs—involves their interaction in a specific way with large molecules —proteins in general . . ."

2. Method, Results and Discussion

A. IONIZATION AND PHARMACOLOGICAL ACTIVITY

In a careful and systematic paper, Lewis (1954) determined the dissociation of a group of sympathomimetic amines, using in some cases both potentiometric and spectrophotometric methods; he tabulated them for two types of pharmacological assay, the contraction of isolated rabbit uterus, and the relaxation of isolated rabbit intestine, the potency of L-epinephrine being expressed as 100. If pK_a and pH are known, the percentage of ionization can be calculated (Albert, 1952). Lewis (1954) found that at physiological pH, the amines existed as 94% cation, 5% zwitterions, and 1% anion plus undissociated molecules. He concluded that "ionization cannot account for the differences in activity among sympathomimetic amines derived from phenylethylamine. Further, at physiological pH the cationic form of these compounds is by far in excess of any other ionic species and is thus probably responsible for pharmacological activity, but other possibilities cannot be excluded."

The compounds investigated by Lewis (1954), with the numbers assigned to them in Figs. 1 and 2, were

1. β-Phenylethylamine
2. (±)-Phenylisopropylamine
3. WIN 5523
4. (−) Ephedrine
5. (−) Propadrine
6. Tyramine
7. WIN 5565
8. WIN 5512
9. (±) *p*-Sympatol
10. WIN 833
11. WIN 5513
12. (−) *m*-Sympatol
13. Hydroxytyramine
14. Epinine
15. (−) Noradrenaline
16. (−) Adrenaline
17. (−) Isopropylnoradrenaline
18. WIN 5589
19. (−) Corbasil
20. WIN 3243
21. WIN 5514
22. WIN 5505
23. WIN 5503
24. WIN 5579

His data were retabulated and, since the overall range in activity of the amines was of the order of 10,000 (4 log units), they were plotted as the

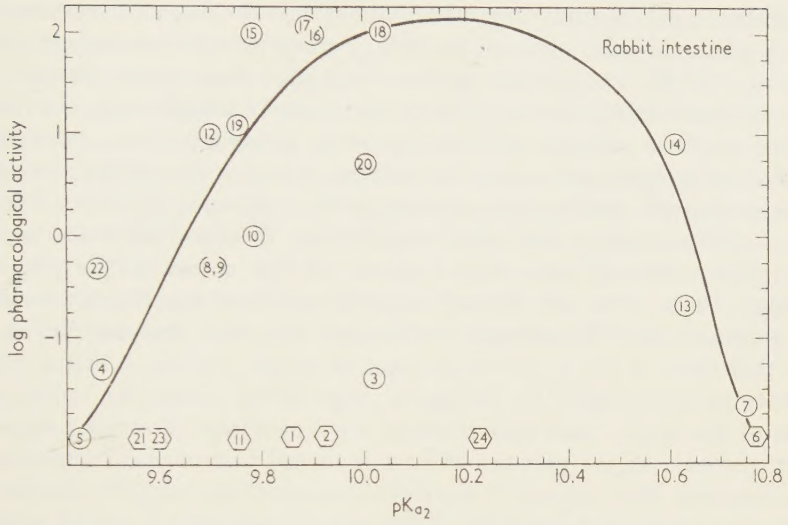

FIG. 1. Functional relationship suggested between pK$_{a_2}$ and pharmacological activity on rabbit intestine, of 24 sympathomimetic amines. Compounds with inappreciable activity (<0·01) shown as hexagons along abscissa. For the circles, between pK$_{a_2}$ 9·44 and 10·03, assuming the linear relation: log pharmacol. activity = 3·29 pK$_{a_2}$ − 31·72, the coefficient of correlation = + 0·54.

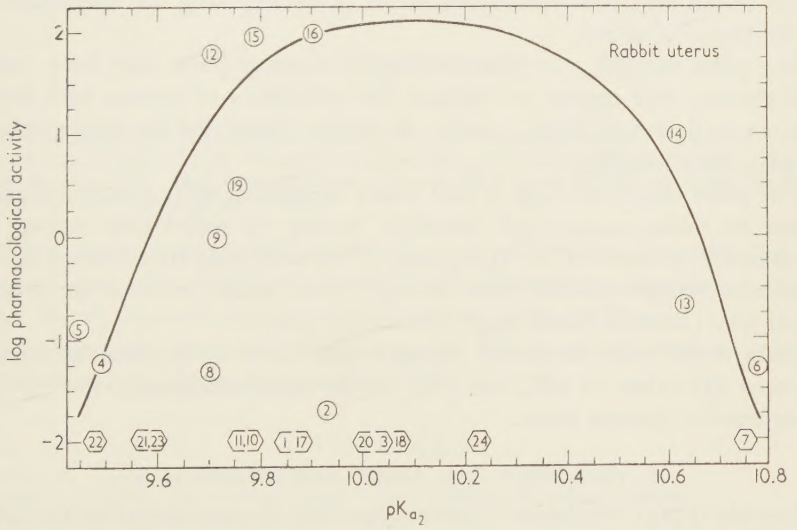

FIG. 2. Functional relationship suggested between pK$_{a_2}$ and pharmacological activity on rabbit uterus, of 24 sympathomimetic amines. For the circles, between pK$_{a_2}$ 9·44 and 9·93, the coefficient of correlation = + 0·38. Compounds with inappreciable activity (<0·01) shown as hexagons along the abscissa.

logarithm of the pharmacological activity against the dissociation constants of the phenolic (pK_{a_1}) and amino (pK_{a_2}) group of the sympathomimetic amines. All the compounds without exception have been plotted. To differentiate readily the active from the inactive compounds, the latter *having no finite activity, and designated by Lewis as* $<0{\cdot}01$, have been plotted as hexagons at -2 on the ordinate, whereas the circles pertain to *those compounds showing finite activity, over a scale of* 4 *log units.* Beyond this, no differentiation was made amongst them. Curves with broad maxima were drawn through the plotted points of Fig. 1 and 2, for pK_{a_2} *vs.* activity. These plots are deemed to be of statistical significance because the points do not fall randomly on the grid. The chief weakness lies in the fact that there is not a continuous spread of pK_a values, a hiatus being evident in mid-range. The hexagons range across almost the entire pK_a scale of the curve. Had they induced weak activities of varying degrees, their positions on the plot would have indicated lack of effect attributable to ionization. Here structure overrides ionization in a negative fashion.

If one assents to the validity of the curves as drawn, it should indicate (*a*) that ionization does play a role, and (*b*) for the synthesis of active compounds by the organic chemist, it would be desirable to search for analogues with pK_a in that region where the function maximizes. The cellular action of the amines appears to be both at the cell surface, as well as within the cell. In the latter case, permeability of the cell surface to the amine would also be a controlling factor in determining the magnitude of the biochemical effect.

The plots for pK_{a_1} *vs.* pharmacological activity have also been made (not shown), and appear to indicate the possibility of curves with broad maxima as drawn in Figs. 1 and 2. However, fewer values were available for pK_{a_1} than for pK_{a_2}.

The plots shown in Figs. 1 and 2 are concerned with pharmacological action on rabbit uterus and intestine. It may be noted that one of the compounds, tyramine (No. 6), appears to act indirectly by releasing epinephrine or norepinephrine from storage sites within or near the arterial vessel wall (Burn & Rand, 1958; Nasmyth, 1960).

From a plot (not shown) it appears that there is no close correlation between the values of pK_{a_1} and pK_{a_2} in the same compound, merely that a positive correlation exists.

B. IONISATION AND AMINE OXIDASE ACTIVITY

Lewis's (1954) tabulation contains also data for the oxidations by amine oxidase, an enzyme widely distributed in tissues (Blaschko, 1952 1957; Blaschko, Richter & Schlossman, 1937). He concluded that no relation exists between enzymatic oxidation and pK_a. However, these too appear

to show a correlation with pK_{a_2}, but of a type different from that shown in the previous curves. Figure 3 seems to indicate a positive correlation between increasing values of pK_{a_2} (amino group) and the relative oxidative rate by amine oxidase. Few values for pK_{a_1} are available, and these do not plot out either linearly or as the concave-upward curves. It is of interest that the curve for the amine oxidase activity (Fig. 3) is different from those for pharmacological activity (Figs. 1 and 2), for the chemical action involved probably is different from that going on during the pharmacological action by the amines on tissues. Blaschko (1952) has assumed that in the oxidative deamination catalyzed by amine oxidase, it is the ionized form of the amine that reacts with the enzyme;

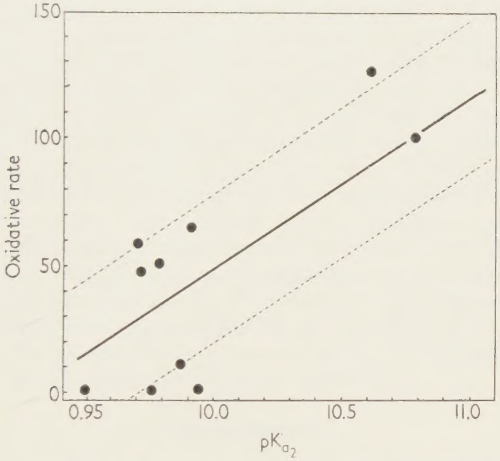

FIG. 3. Suggested relationship between pK_{a_2} of sympathomimetic amines nos. 1, 2, 4, 6, 9, 12, 14, 15, 16 and 19, and enzymatic oxidation by amine oxidase. Solid line fitted by least squares. The dash lines represent the standard error of the estimating equation: oxidative rate = 66·9 pK_{a_2} − 619·5.

he cites experimental evidence for this. At pH values favorable for the enzyme's activity, those amines which are substrates are present mainly as ions. Poorly ionized amines were not appreciably oxidized.

The various chemical mechanisms by which catecholamines may be inactivated have been examined by Zeller (1959).

REFERENCES

ALBERT, A. (1952). *Pharmacol. Rev.* **4**, 136.
ALBERT, A. (1960). "Selectivity Toxity", 2nd Edition. Methuen, London.
BLASCHKO, H. (1952). *Pharmacol. Rev.* **4**, 415.
BLASCHKO, H. (1957). In "Psychotropic Drugs". (Garattini, S. and Ghetti, V., eds.) Elsevier, Amsterdam, London, New York and Princeton.
BLASCHKO, H., RICHTER, D. & SCHLOSSMAN, H. (1937). *Biochem. J.* **31**, 2187.
BURN, H. H. & RAND, M. J. (1958). *J. Physiol.* **144**, 314.
DAVSON, H. (1951). "A Textbook of General Physiology", p. 187. Blakiston, Philadelphia.

DAVSON, H. & DANIELLI, J. F. (1952). "The Permeability of Natural Membranes", 2nd Edition. Cambridge University Press, London.
HÖBER, R. *et al.* (1945). "Physical Chemistry of Cells and Tissues". Blakiston, Philadelphia.
IRWIN, M. (1926). *J. gen. Physiol.* **9**, 561.
LEWIS, G. P. (1954). *Brit. J. Pharmacol.* **9**, 488.
NASMYTH, P. A. (1960). *J. Physiol.* **152**, 71P.
PAULING, L. (1959). In "Significant Trends in Medical Research", p. 13. (Wolstenholme, G. E. W., O'Connor, C. M. & O'Connor, M., Eds.). Little Brown, Boston, Mass.
SCARTH, G. W. & LLOYD, F. E. (1930). "An Elementary Course in General Physiology". Wiley, New York.
WEST, E. S. (1956). "Textbook of Physical Chemistry". Macmillan, New York.
ZELLER, E. A. (1959). *Pharmacol. Rev.* **11**, 387.

J. Theoret. Biol. (1961) **1**, 295–317

The Logical Analysis of Animal Communication

PETER MARLER

University of California, Department of Zoology, Berkeley, U.S.A.

(*Received* 16 *January* 1961)

An attempt has been made to describe some of the responses evoked by communication signals in certain animals and to infer the kind of information which the signals transmit. Using the methods developed by C. W. Morris (1946) for the logical analysis of human language, identifiors, designators, appraisors and prescriptors can be distinguished. Animal signals are rich in designative information, and five subcategories are distinguished: species-specific, sexual, individual, motivational and environmental information. The influence of natural selection upon the form of a signal will vary according to its information content. For example, the variable nature of some signals and the stereotypy of others can be related to the conveyance of different types of motivational information. A single signal often conveys several different items of information which are usually inherent in the whole signal and not represented by different parts of the signal. The form of some signals is arbitrary but the physical structure is often directly related to information content, in an iconic manner, or in other ways.

Introduction

By any reasonable definition of the term "communication" there can be no doubt that animals communicate with each other. Some authors even extend the term to include exchange of stimuli between organisms and their physical environment (Stevens, 1950), which is perhaps further than it is necessary to go. The position adopted in a recent book by C. Cherry (1957) serves very well to restrict the discussion to a social context. He defines communication as: "The establishment of a social unit from individuals by the use of language or signs". Inclusion of both signs and language in this definition ensures from the outset that studies of communication systems shall not be restricted to the languages of man. This simple step, which so many past authors have been reluctant to take leads Cherry into a lucid, illuminating account of the properties of communication systems and of the methodological problems which they pose. As a student of animal behavior who has been grappling with problems of animal communication the writer has been struck by the relevance to zoology of many of the ideas expressed in Cherry's book. This paper tries

to apply some of them to animal communication and to show that they can open up new avenues to the understanding of the kind of evolutionary problems with which many zoologists are concerned.

THE ANTHROPOCENTRIC APPROACH

Comparative psychologists have neglected the subject of animal communication to a remarkable degree—remarkable, that is, until one reflects on the anthropocentric point of view of most psychologists. The strictures imposed by F. A. Beach (1959) on comparative psychology are nowhere more relevant than in the subject of animal communication. The main concern has been to differentiate man and the animals, rather than to determine the properties which their "languages" may have in common. Dozens of cases could be cited where this prejudice has influenced the questions that are asked, and therefore the answers obtained. Even in such distinguished contributions as the chapter on the social significance of animal studies by D. O. Hebb & W. R. Thompson (1954) in the "Handbook of Social Psychology" this bias is evident. After discussing the human capacity to combine and readily recombine sounds for different effects, they acknowledge that language has other distinctive characteristics but assert their belief that the above criteria "are enough to set it off fully from animal communication". A promising discussion thus terminates at the point where it is about to become productive. In the field of linguistics attempts to analyse animal communication have also been marred by anthropocentric viewpoints (e.g. Revesz, 1956), although it is also a linguist, C. F. Hockett (1961), who has succeeded in defining the properties of human language in a manner that permits us to test for their occurrence in animals. In doing so, he has omitted "purposiveness" as one of the criteria. This concept, which may also be associated with an anthropocentric viewpoint, has bedevilled investigations of animal behavior in the past (Thorpe, 1956).

PURPOSIVENESS

Hebb and Thompson (1954) question whether the waggle dance of the honey bee (von Frisch, 1954) is purposive, suggesting that it would be if:

(*a*) only the first of several returning bees made the waggle, since, if the message has already been conveyed to the colony by ten other bees, there is little sign of purpose in behavior that conveys it once more; and

(*b*) the worker still made the dance as though the audience was present even when it had been removed.

(*a*) seems to be based on the misconception that the entire contingent of perhaps ten thousand workers can perceive the performances of a dozen

or so dancers. The solution to (*b*) is not certain, but personal observations suggest that an eager audience in the hive is certainly stimulating to a dancer. However, we may ask whether this is a sign of purposiveness, or whether the dancer is simply stimulated through palpation by the antennae of the audience.

If the concept of purposiveness has to be reduced to such a vague level before it can be tested, as Hebb and Thompson seem to imply, we may wonder whether it has not ceased to be valuable as a theoretical construct in the analysis of animal behavior. W. H. Thorpe (1956) has pointed out how difficult the subjective and objective aspects of purposiveness are to separate. It may be best to restrict the idea of purposiveness to a human context. Hebb & Thompson (1954) state that the essence of purposive communication is that "the sender remains sensitive to the receiver's responsiveness during sending, and by modification of his sending shows that his behavior is in fact guided by the intention of achieving a particular behavioral effect in the receiver". By this definition any dog-fight qualifies as purposive, as the authors admit. It is not clear what is gained by using a specialized and loaded term for a process which is basically a mutual communicatory exchange, unless to draw attention to possible subjective phenomena. If the latter, then we should recall Thorndike's (1911) still relevant warning about the dangers of the introspective method in animal studies, notwithstanding Tolman's (1932) demonstration that by placing special interpretations upon it, purposiveness can be given an objective basis.

AN OBJECTIVE APPROACH

The descriptive or taxonomic approach, which comes less readily to psychologists than to zoologists, has provided the bulk of our present knowledge about animal communication, as applied by such classic investigators as Charles Darwin and K. von Frisch. This in turn has led to new inductive generalizations by K. Z. Lorenz, N. Tinbergen, W. H. Thorpe and others which will provide the framework of future work for many years to come. Instead of approaching animal communication with anthropocentric preconceptions, they set out to describe the natural behavior in objective terms, seeking to derive conclusions about the evolutionary basis of behavior. Even such severe critics as D. S. Lehrman (1953) fully acknowledge the great value of the advances which this "ethological" school has achieved. Communicatory behavior has figured prominently in this work and provided the basis for much of the theoretical discussion in the early papers of Lorenz (1935) and Tinbergen (1940). The scope has subsequently been broadened to include other types of behavior, and the "ethological" school (Tinbergen, 1951; Thorpe, 1956) now provides a rationale for the analysis of animal behavior.

In proceeding thus far, it is the author's contention that some of the special circumstances surrounding communicatory behavior have been overlooked. Close attention has been given to the evolutionary basis of visual signals and the motivation which underlies them. Less attention has been given to the nature of the actual communicatory process; to the questions raised by the process of exchange of signals between one animal and another. The psychologists' concern with this aspect leads them to a consideration of purposiveness, but this does not prove to be a productive line of attack. A strictly objective approach is required which can be applied with equal efficacy to the communication of animals and of man. This paper seeks to show that the theoretical framework presented by Cherry (1957), building especially upon the ideas of Pierce (see Gallie, 1952) and Morris (1946), provides us with such an approach which can lead to advances in our understanding of animal communication.

Semiotic: The Theory of Signs

Dissatisfaction with the results of previous attempts to separate the subjective and objective aspects of human language led C. K. Ogden and I. A. Richards, in a book called "The Meaning of Meaning" (1923), to consider the implications of the theory of signs (or symbolism as they sometimes call it) as developed by the logician, C. S. Pierce. The relationship between a word or symbol and its external referent is shown to be elusive. Perception of external objects (referents) always involves sign situations. We respond only to a part of the whole object. That part comes to represent the whole object as a kind of symbol or sign. "If we realize that in *all* perception, as distinct from mere awareness, sign situations are involved we shall have a new method of approaching problems where a verbal deadlock seems to have arisen. Whenever we perceive what we name a chair we are interpreting a certain group of data, and treating them as signs of a referent." Narrowing down the discussion to the use of language they suggest that "when we consider the various kinds of sign situations . . . we find that these signs which men use to communicate with each other and as instruments of thought occupy a peculiar place". This comes to bear directly on our present problem with the statement that "the person actually interpreting a sign is not well placed to observe what is happening. We should develop our theory of signs from observations of other people, and only admit evidence from introspection when we know how to appraise it."

To explain the approach of C. S. Pierce to the problem of language analysis, W. B. Gallie (1952) gives the following example. "Suppose that in any particular case we are in doubt whether some sign made by an

individual A has been interpreted or understood by a second individual B. How should we set about trying to settle the question? Should we somehow or other try to discover directly what B's 'mental reaction' has been? It seems quite certain that we have no means whatever of doing this. What we would do, surely, is to try to discover whether B has made some overt response such as A's sign would justify." Cherry (1957) emphasizes the same point, that only a non-participant observer can make fully objective observations on communication systems.

The science of semiotic has arisen to deal with the kind of data that are obtained by direct, non-participant observation of communication systems. It is usually divided into three parts: *syntactics*, the formal study of signals as physical phenomena, and the laws relating to them; *semantics*, study of the "meaning" of signs; and *pragmatics*, the significance of signals to the communicants (Cherry, 1957). The application of syntactics to animal communication is clear, and great progress has been made by Tinbergen and others in this kind of analysis, especially in the sphere of visual communication (see Tinbergen, 1940, 1951, 1952, 1959). Semantics are of doubtful value in animal studies, and as Cherry points out there is considerable overlap with pragmatics, even in the sphere of human language. Pragmatics on the other hand forms the natural complement to syntactics, one defining the physical properties of signals, the other concerning itself with the role of those signals in the communicatory process, a role which we seek to establish by observing and interpreting the response which they evoke in other animals.

ANIMAL PRAGMATICS

The central problem is to determine the nature of the information content of communication signals. As Cherry points out "information content is not to be regarded as a commodity; it is more a property or potential". It cannot be discussed independently from the occurrences of responses to the signal in other organisms. We thus require a means of inferring information content from the nature of the response given. We may note in passing that the information theory developed by Wiener & Shannon (Shannon & Weaver, 1949) is of no help to us here since it operates only "at the syntactical level" (Cherry, 1957). The work of C. W. Morris (1946), however, is directly concerned with analysis of human language at the pragmatic level and can give us some clues as to how to proceed.

Morris seeks to distinguish between signals which function as *identifiors*, *designators*, *appraisors* and *prescriptors*. He emphasizes that this is not an exhaustive list, and elaborates some of them further to deal with special problems of human language. The four basic categories will suffice as a

basis for further discussion. We can describe each of them as conveying a corresponding type of information, provided that we can discern an appropriate response from a communicant. The categories are not mutually exclusive, so that one signal might convey one or all of the different types of information.

Morris defines the four categories as follows: "In the case of *identifiors*, the interpreter is disposed to direct his responses to a certain spatio-temporal region; in the case of *designators* the interpreter is disposed toward response sequences which would be terminated by an object with certain characteristics; in the case of *appraisors* the interpreter is disposed to respond preferentially with respect to certain objects" as manifest in a choice situation; "in the case of *prescriptors*, the interpreter is disposed to perform certain response sequences rather than others." So identifiors may be said to signify (i.e. convey information about) location in space and time, designators to signify characteristics of the environment, appraisors to signify preferential status and prescriptors to signify that specific responses are required. This classification cuts across the division of language into emotive and referential (or symbolic) which received so much emphasis from Ogden & Richards (1923). Morris shows how his classification is subject to testing in a way that the other is not. Moreover we can see that while prescriptors and appraisors embody much of the quality of "emotive" language, and identifiors and designators are more obviously "referential"; in nature, the latter can be emotive in certain circumstances. Thus the new approach is more precise and should be regarded as replacing the older terminology, as Morris suggests.

We now have to demonstrate that this method of analysis can in fact be applied to animal communication systems. J. B. S. Haldane (1953; Haldane & Spurway, 1954) have already shown some ways in which this may be done, and the writer also made an attempt to analyse vocal communication in a small bird, the chaffinch (Marler, 1956) by a method similar to the one suggested here. A reinterpretation of those same data can serve as an illustration. In essence, given a knowledge of the response of other animals to the signal and of the other circumstances in which that same response is given, we can infer the nature of the "message" transmitted by the signal.

The song of the chaffinch is given only by the male. The species is normally monogamous, and the song is especially frequent in an unmated male, given only within his territory. An unmated female chaffinch in reproductive condition responds to repeated singing by persistently approaching the singing male, soliciting his courtship, and eventually establishing a pair bond with him. Circumstantial evidence suggests that some females learn the individual characteristics of their mates' song, and

subsequently respond to them in a preferential way. The behavioral exchanges consequent upon the female's response to the song are confined to a sexual context and are normally evoked by what we may describe as an "appropriate sexual partner". We may infer that frequent male singing conveys information about this particular class of objects which are the "designata" of the male's song, in this situation. What exactly is the information content which is implied?

An appropriate sexual partner for an unmated female chaffinch in reproductive condition is *an unmated male chaffinch in reproductive condition, in possession of a territory* (within which nesting will take place), *who is close to a location occupied by the female at the same time as she is there*. We are suggesting that all of these items of information are conveyed to her by the male's song. This does not imply that the song has any meaning for her, only that it performs selective actions upon her, appropriate to a certain input of information (Cherry, 1957). The male's individual identity may also be conveyed in some cases. To what extent can this be fitted into Morris's scheme?

"Identifiors" dispose the receiver to direct his responses to a certain spatio-temporal region. We can show that such identifying information is present in the male chaffinch's song which provides an abundance of clues for precise location of the singer in time and space (Marler, 1959). In some respects "locating" information might be a better description.

"Designators" dispose the receiver towards response sequences which would be terminated by an object with certain characteristics. Designative information is thus to be defined by the characteristics of the object normally evoking the response, in this case those of an appropriate sexual partner. This would encompass all of the items outlined above and we shall suggest in a moment that further sub-categories may be desirable.

"Prescriptors" dispose the receiver to perform certain response sequences rather than others. The response prescribed for the female chaffinch is to approach and to adopt certain postures which elicit male courtship. Prescriptors and designators may be confused in some cases because we need to know the kind of response prescribed before the object designated can be discovered. Circular reasoning can only be avoided when prescriptor and designator are contained in different signals. If they can be combined with other signals a different response can be prescribed with the same designator and the effects can be separated. When the same signal performs both functions, as seems to be common in animals, no logical separation between prescriptors and designators is possible.

Appraisors dispose the receiver to respond preferentially to certain objects. Although we have no quantitative information, the frequency with which a song is repeated probably conveys such appraisive informa-

tion. Within the range of song frequencies that will evoke a response, a female confronted with two singing males may be most likely to choose the one who is singing most persistently.

A more detailed breakdown of the nature of designative (and therefore prescriptive) information is required if this system is to aid us in analysis of the evolution of animal communication systems. Most critical from the point of view of natural selection is the presence of the *species-specific* information—that the singer is a chaffinch. We can also separate *sexual* information—that the singer is a male; *individual* information—that the singer is a particular individual; *motivational* information—that the singer is in reproductive condition; *environmental* information—that he is within his territory and has no mate. The criteria by which these types of designative information may be identified are as follows.

Species-specific information and its evolutionary implications

If the response given to the signal is normally evoked only by members of one species we may infer that species-specific information is conveyed by the signal. Usually a member of the same species will be involved, since many animal communication signals play a role in reproductive isolation. Information about other species could come into this category, as for example in the signals exchanged between a commensal and its host. There are also mimics which emit signals with a false species specificity.

Some signals are lacking in such species-specific information. For example, in a situation involving acute danger male chaffinches have an alarm call consisting of a high thin squeak. It is typically given in response to a hawk flying overhead. It evokes the same response from other chaffinches as the stimulus provided by the hawk, namely, direct rapid flight to the nearest cover. However, several other small woodland birds have converged upon the same type of alarm call presumably because, as mentioned below, it is a difficult sound to locate, and so exposes the caller to a minimum of danger. Chaffinches will respond to the corresponding alarm calls of other species as promptly as to their own. Such cases of interspecific communication are very common in the woods in which chaffinches live. Thus species-specific information is not present in this call. Degrees of species specificity may be expected, decreasing to the extent that signals are of mutual value in communication within a group of different sympatric species.

A signal functioning to transmit species-specific information will be subject to certain evolutionary pressures, since there must be a minimum of confusion with signals used by other species at the same time and place. Circumstantial evidence suggests that many auditory and visual signals have been selected for specific distinctiveness (see Sibley, 1957). Con-

versely signals with an interspecific function may be subject to selection for convergence upon a common type—or at least to a minimum of selection for divergence. Where species specificity is required, it is desirable that, as well as being specifically distinct, the signal should also be biologically improbable and conspicuous for effective communication against a background of environmental "noise" (Lorenz, 1951). A relative lack of variability is also required among members of the same species, or at any rate of the same population, an important point when we compare signals which convey individual information.

Sexual information. Responses associated with reproduction are normally evoked by members of the opposite sex when in the appropriate physiological condition. A signal evoking such a response may be said to convey sexual information. There are, however, cases where such behavior patterns are also evoked by members of the same sex in what may be called homosexual or pseudo-sexual behavior (see Morris, 1955). The incidence of sexual information varies considerably as manifest in the extent to which the sets of communication signals of the male and female overlap in different species. The same principles often apply to visual and auditory signals, so that the more sexually dimorphic finches, for example, also show the greatest discrepancy between the repertoires of displays and vocalizations in the two sexes (Hinde 1955–6). The principles governing these variations in the prevalence of sexual information in the signals of different species have not yet been worked out.

In discussing differences between the signals produced by male and female animals, Hockett (1961) has elevated the principle of what he calls "interchangeability" to the level of a major criterion in the analysis of communication systems. He suggests that while it occurs in animals, it is especially characteristic of human language, implying that any person can theoretically reproduce sounds made by any other person. He makes a distinction between language and paralanguage (Trager, 1958) and applies the principle of interchangeability particularly to the former. However the same distinction, which seems to rest on an intuitive judgement with reference to human language, cannot be made with animals. If we regard the difference between the sexes as a means of conveying sexual information, this information is obviously present as a conspicuous and more or less consistent difference in frequency between the speech of men and women. While the auditory signals produced by women share many characteristics with the corresponding signals of men, there are also in Western Society certain unavoidable differences of pitch, unavoidable, that is, for *most* women (Potter, Kopp & Green, 1947). In this respect the lack of "interchangeability" in human speech is more striking than in some animals, since even strongly sexually dimorphic species often have

some signals which are consistently identical in all respects in the two sexes.

Individual information. The transfer of individual information by a signal is implied whenever the response is normally only evoked, or most readily evoked, by the particular individual emitting the signal. The qualifications admit the possibility of appraisive information being included here, since the female chaffinch, for example, will respond to an unfamiliar chaffinch song, though she may choose a familiar song if given a choice. In many circumstances individual recognition of the signals of mates, rivals, young, and companions plays an important role in the social behavior of animals (Nice, 1943; Marler, in press).

A signal which transmits individual information is subject to selective influences different from those associated with species-specific information. The latter, as we have seen, is most readily transmitted by signals which show little variation, either in the individual or within a population of a given species. Individual information is again best conveyed by signals which vary little in the individual. But it is also a prerequisite that the signals emitted by individuals of the same species, especially within the same population, should differ from each other in a consistent manner. Circumstantial evidence suggests that there is an unduly high degree of intra-group variability in signals which are thought to be involved in individual recognition, such as visual signals originating from the head region of birds, and the songs of some species of birds (Marler, 1959, in press). Some bird songs appear to convey both species-specific and individual information by relegating the stereotyped and variable properties to different parameters of the song.

Motivational information. The last two categories of designative information, motivational and environmental, are the most difficult to define, the least understood, and perhaps ultimately the most important from an evolutionary point of view. The transmission of motivational information by a signal may be inferred if the response given is appropriate to a particular motivational state of the signaller. Such a signal conveys information about variations in the readiness of the signaller to engage in certain classes of activity, such as feeding, fighting or copulation and so on.

The male chaffinch's song evidently communicates to the female the fact that he is in a reproductive state. This condition usually lasts for about three or four months. Short-term changes in motivation may also be communicated by signals. When a mated female has built a nest and is preparing to ovulate she will allow the male to copulate at intervals for about four or five days. When actually ready for copulation she gives a special call which is restricted to this context. The male promptly

approaches and mounts. Similarly the calls given periodically by the young as they become hungry, cause the parents to bring food to them.

Information about still more subtle changes in motivation can also be transmitted. Here the best evidence comes from visual signals, and to discuss them we shall again have to anticipate consideration of the divisible parts of the signal and the information they convey. Many of the communication signals used by animals are subject to what Morris (1957) has called the "principle of typical intensity". This implies that the signal varies little or not at all, with variation in the level of motivation with which it is associated. Either it is given in "typical intensity" or it is not given. Such a signal can effectively communicate presence or absence of a certain type of motivational information but not variations in degree. For many purposes this appears to suffice. In general, a male chaffinch is either in reproductive condition or he is not, and an "all-or-none" type of signal can communicate this.

Other signals do not obey the principle of typical intensity, but vary widely in form, completeness and frequency with the intensity degree of motivation with which they are associated. Visual signals used in fighting behavior are particularly prone to vary in form with slight variations in the presumed balance between the tendencies to attack and withdraw. An opponent is often highly responsive to the slight shifts in motivation which these changes convey, advancing in response to signs of withdrawal, and vice versa, and the final outcome of the fight will normally be determined in this way. On the basis of his extensive studies of the behavior of cats, Leyhausen (1956) has been able to construct a Latin square of the changes in facial expression with changes in aggressiveness and readiness to flee, including all possible combinations between the two, a remarkable demonstration of the complex array of motivational information that such graded communication signals could convey. A function of this kind obviously has profound effects upon the way in which the signals will evolve.

The signals discussed above convey what we can describe as "positive" motivational information; they enable a receiver to "make a positive prediction" of the response which the signaller is likely to give when approached. The evolution of a second class of signals has been governed by a trend towards becoming the direct opposite of other signals, as Darwin (1872) pointed out with his principle of antithesis. His classical example is the behavior of a submissive dog which can only be described as the opposite in all respects of a dog which is fighting. Many other examples of such "antithetic" or "reversed" signals (Tinbergen, 1959) have been described, having the function of conveying something like "no offense meant", and so reducing the chance of an open conflict occurring (Tinbergen & Moynihan, 1952).

In the light of the present analysis we can reinterpret this function as the conveyance of negative motivational information, making it possible for the receiver to predict that the signaller will *not* behave in a certain way when he is approached. All of the cases known so far occur in potentially aggressive situations and appear to function by reducing the chances of attack or flight, or both. Negative information about readiness to attack or to flee is conveyed in most cases. Once again there are evolu-

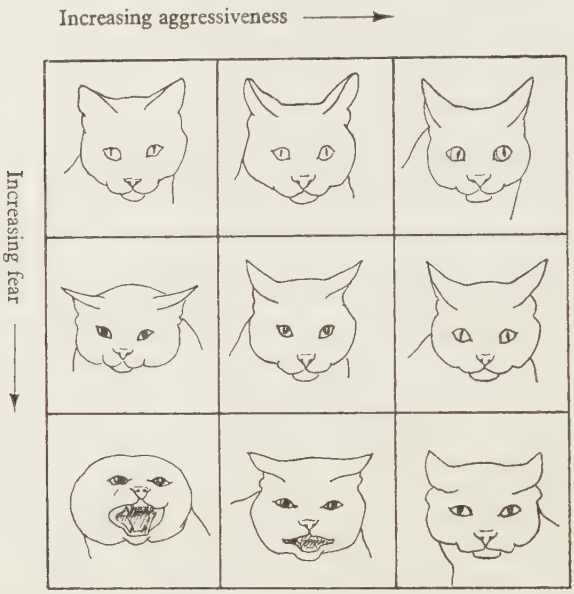

FIG. 1. Changes in the facial expressions of cats associated with variations in the intensity of aggressiveness and fear. (After Leyhausen, 1956.)

tionary implications which could be explored further. For example, aggressive displays usually have certain formal properties and also a certain orientation with respect to the opponent. A negative element can be introduced with respect to a certain receiver, both by reversing formal elements of the display and also by orienting away from the particular receiver. Both trends, in various combinations, can be traced in the examples given by Tinbergen (1959), ranging from simple reorientation of an aggressive display to the reversal of other aspects as well.

Environmental information. When Morris (1946) set up the category of designators he visualized their primary role as conveyors of environmental information, encompassing as they do the characteristically human tendency to give things names. In animals we may infer that a signal has

transmitted environmental information if the response it evokes is appropriate to some characteristic of the environment of the signaller at the moment or in the immediate past. The exact temporal relationship will be discussed in a later section. If, for example, a particular sound is produced in the presence of food, as occurs in herring gulls, and if others respond by approaching and looking for food, as Frings, Frings, Cox & Peissner (1955) have shown to be the case, we may infer that information about the presence of food was conveyed by the signal.

We have inferred that the male chaffinch's song conveys two items of environmental information, one positive, that he is within his territory, and one negative, that he has no mate. Other examples are mainly concerned with what are perhaps the two most important aspects of the physical environment to animals, food and predators. The use of signals conveying the presence of food is probably widespread within family groups. The mock pecking movements, by which a domestic chicken attracts her chicks to a supply of grain, are a familiar example. Special calls are probably also used in this context, though no examples are known to the author. The gull call mentioned above certainly attracts adult gulls as well as young (Frings, Frings, Cox & Peissner, 1955).

The best known signals concerned with the communication of food are the dances of honey bees, analysed in detail by Karl von Frisch (1954). The round dance communicates the distance of a food source in the neighborhood of the hive and its richness. The waggle dance, when given in the hive, also communicates the direction, as well as the distance and richness of sources which are further from the hive. This is, in a sense, another case of communication within a family group, for the worker bees are all daughters of the same queen. The dances are also used in the swarm, to communicate the distance, direction and suitability of new home sites (Lindauer, 1957). Here the dances are given, not on the vertical combs in the darkness of the hive, as in the case with the food dances, but on the surface of the swarm. Clearly the context in which the dance is given affects its communicatory significance. In both cases the signal is a mechanical one, received particularly by the bee's antenna and the mechanoreceptors at its base. The different items of information are conveyed by different aspects of the dance: direction by the angle of the waggle run with respect to the vertical; distance by the tempo of the dance; richness of the food source by the persistence of the dancing. Information about the latter might also be placed under the heading of appraisive information, since it determines the choice made between different situations, particularly in the swarm, where departure from the temporary resting place does not occur until the dancers have reached a degree of unanimity. The final decision is achieved by the scout bees discovering the best site who, by

their more persistent dancing, eventually sway those who have discovered alternative but less satisfactory sites (Lindauer, 1957).

If we place the environmental information conveyed by animal signals on a specificity-generality continuum it usually appears to be relatively unspecific in nature. Information is transmitted about food but not about which particular food. It is true that foraging honey bees may pick up the scent of flowers and so convey the identity of the nectar source to members of the hive (von Frisch, 1954), but the specificity of the signal has been evolved by the plant rather than by the bee. Similarly signals concerned with communicating danger usually seem to do little more than signify different degrees of danger without specifying which environmental agent is responsible. Many small birds have different vocal signals for sudden, acute danger, such as when a hawk appears overhead, and for less dangerous situations, such as when they discover a sleeping owl, or a cat on the ground beneath them. The responses which these stimuli evoke are quite different, sudden flight to cover and cryptic behavior on the one hand; approach to a safe distance and conspicuous "mobbing" behavior on the other. However, the circumstances may greatly modify the response to a given predator. In early spring a male chaffinch will normally give the call associated with acute danger only in response to a flying hawk. But later when he has nestlings, a variety of animals will evoke this call if they come near the nest. Thus we cannot say that the call communicates the presence of a hawk. There may be animals which convey more specific environmental information. Our knowledge is so fragmentary that we cannot begin to generalize. The European willow warbler is thought to have two mobbing calls, one given to a perched hawk, the other to a cuckoo, suggesting that more specific information may be conveyed in this case (Smith & Hosking, 1955).

Conclusions on information content. Although identifiors (or locators) and appraisors occur among the communication signals of animals, designators seem to be most richly represented. Five different categories of designative information have been described, all having particular implications for the evolution of animal communication systems. Perhaps the most prominent category in human language is environmental information, the one to which investigators most often turn when they wish to compare the animals with man. The basic capacity, to convey environmental information by signals, is present in both. However, the time element in this process is significant as several authors have pointed out (Haldane, 1956; Haldane & Spurway, 1954; Hockett, 1961). In animals the delay between perception of an object in the environment and emission of a signal conveying information about that object is usually a short one. In man the delay may be extended almost indefinitely, illustrating what Hockett

(1957) calls "displacement". The only well documented case involving a longer delay in animals comes again from the honey bee where the dance occurs after the forager has returned to the hive (von Frisch, 1954). Here there are finite limits to the delay, which is short by human standards in any case, and it would hardly be useful to the honey bee if it were any longer since the food supply from a given plant varies from hour to hour. Human capacities in this direction are probably unique, although one may wonder if any but the most educated observer would be able to detect such extensive time delays in animals even if they occurred.

The context may have considerable significance to the animals themselves. Hockett (1961) has pointed out how the responses of honey bees to dancing differs when it takes place in the hive and on the swarm. In speaking of the communicatory process as though it were mediated by signals alone, we have thus been guilty of over-simplification. The response evoked by a signal—and therefore the information it conveys—may vary with changes in the circumstances both of the sender and the receiver. The song of a male chaffinch is seen in a different light if we observe the response of male chaffinches instead of females. A male chaffinch intruding into another's territory will flee if he hears the owner's song, implying reception of a further item of motivational information, that the owner is ready to attack male chaffinches found within the boundaries of his territory. The response of a male in an adjacent territory will be different again, and so on. The separation of all of the factors which bear on a given act of communication is thus an imposing task. The additional possibility always exists that the signaller may be emitting several different cues at the same time, as seems to be the case in rats, for example, where olfactory, tactile, visual and auditory signals may all play a part in the female's sexual responses to the male (Beach, 1942, 1947).

Divisible Elements of the Signal

In trying to determine the role of prescriptive information in animal signals we have been confronted by the dilemma that it cannot be distinguished from designative information in signals consisting of one indivisible unit. Only when prescriptors exist in physically separate parts of the signal can an unequivocal separation be made. It thus becomes important to transfer our attention from pragmatics to syntactics to consider the physical nature of some of the signals used in animal communication.

CONTINUITY VERSUS DISCRETENESS

Attention has been drawn to the fact that some signals vary to the extent that they sometimes grade continuously into other signals; others tend to appear in an all-or-none fashion, so that they are separate and discrete from

all other signals. The degree of variation observed can be correlated to some extent with the information which the signal conveys. A degree of continuous variation may occur in at least three different circumstances.

First, appraisive information appears to be most commonly conveyed in animals by signal characteristics which vary in a continuous manner. The frequency with which the chaffinch song is repeated probably conveys appraisive information to the female about the male's relative suitability as a mate (cf. page 302). Similarly the persistence of dancing in the honey bee, as expressed by the number of dances given before the sequence is broken, conveys appraisive information about the richness of the food source or the suitability of a new nest site to other members of the hive. Cases may exist where appraisive information is conveyed by a discontinuous series of signals. For example, the remarkable series of postural displays given by the black-headed gull correlated with variations in the relative and absolute levels of tendencies to attack and to flee should come into this category, for while some intergrade, others are discrete, with a sudden switch from one to the other as the balance of motivation shifts (Moynihan, 1955; Tinbergen, 1959). However, it appears that this condition, which is characteristic of human language, is rare in animals.

Another function of continuously variable signals is the conveyance of subtle changes in motivational information. Some signals, as Morris (1957) has indicated, vary little with slight changes in the signaller's motivation, whereas others mirror the changes in motivation very closely (cf. page 308). Particularly in fighting behavior, where the communication of such subtle motivational information can assume great importance, such variable signals are often used. Human language is in some ways less well adapted to convey such continuously variable information because of the tendency to divide continuous phenomena into discrete classes, which is perhaps one of the reasons why animal signals of this type are difficult for zoologists to describe.

Continuously variable signals also occur as a means of conveying environmental information of a continuously variable nature. The best example is again from the honey bee dances, in which both the direction and distance of the food source are communicated. The former is conveyed by the angle of the waggle run with respect to the vertical, the latter by the tempo of the dance, both varying in a continuous manner. No doubt further examples will be discovered.

We may conclude that continuously variable signals have an important role to play in the communication systems of animals. More stereotyped, discrete signals are also common and, for example, make up the bulk of the vocal signals of such birds as the chaffinch. Continuously variable signals have certain disadvantages. Their interpretation may be slow, and

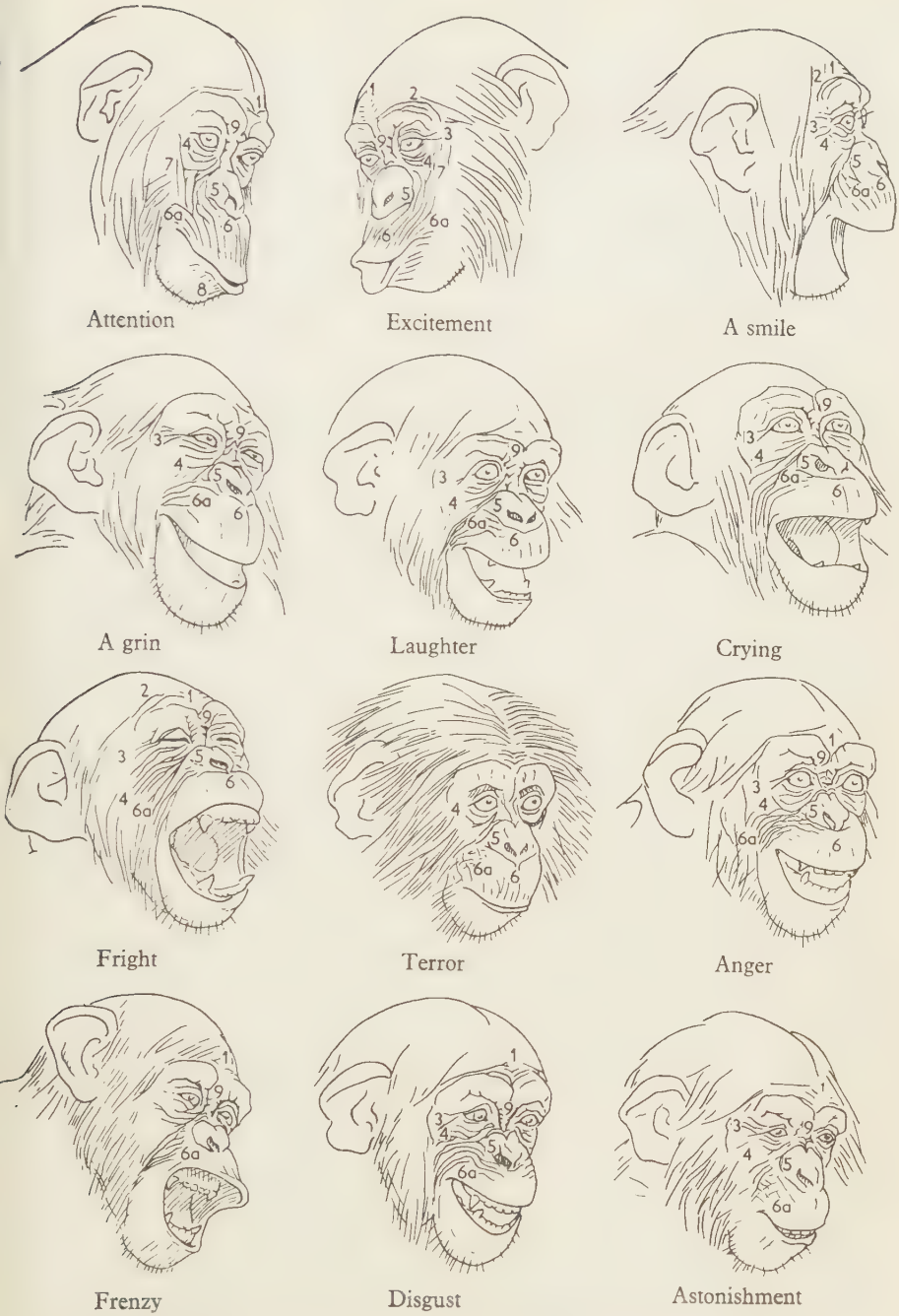

FIG. 2. The facial expressions of a young chimpanzee in various moods. The creases on the face are numbered to emphasize that each one may be involved in several different expressions. (After Kohts, 1935.)

subject to error. Also appropriate inborn responsiveness to all properties of the signals, which characterizes the communication systems of many animals is more easy to visualize with discrete signals than with continuously variable signals. Finally there may be conflicts with other items of information conveyed by the same call. We have seen that the communication of species-specific and individual information both call for stereotyped signals, a requirement which may well override the need to convey subtle changes of motivation.

DIVISIBILITY

The impossibility of separating designative and prescriptive information hinges on the fact that in animal communication systems several items of information seem to be conveyed by one discrete, indivisible signal. We do not normally find the different items of information represented by different elements as is commonly the case in human language, where the component elements can be rearranged to create new "messages". However, this does seem to occur in some cases, particularly in visual communication signals. The facial expressions of chimpanzees probably serve as signals in intraspecific social behavior (Hebb, 1946). In describing them, Kohts (1935) took great care to point out that the same creases on the face may be involved in several different circumstances, the expressions as a whole presumably conveying different information. In her drawings, she even went so far as to number the facial lines to emphasize this point (Fig. 2). Assuming that all elements are necessary for the complete signal (which is difficult to test with visual signals), it appears that the divisible parts can be rearranged to create new "messages". In a similar way the sharing of components by different visual displays of such birds as the chaffinch may imply something similar.

Examples may also be found among vocal signals, but we have to proceed with care. Thus Hockett (1961) quotes Lanyon as presenting evidence that basic motifs in the songs of certain birds are rearranged in different ways to create new songs. A number of cases of this have been described, but there is no evidence that these recombined elements differ in any way in information content. Better examples are likely to be found in the alarm calls of certain birds. Some species have several discrete calls which are given, sometimes alone, sometimes together with other calls. The sequences of different signals may conceivably contribute to one overall signal whose information content varies with the constituents of which it is made up. If this proves to be the case, we are then approaching, at a very primitive level, the kind of lability in the manipulation of the information content of signals which is such a distinctive property of human language.

We must not assume that the lack of such lability among animals is simply a result of incapacity of the nervous system to handle such complex information. The way of life of most animals is so stringent and fraught with dangers that a high premium is placed upon quick production of brief signals, which can be accurately interpreted by receivers, often without the opportunity for previous practice. Given that the fight for survival is controlled by a limited number of factors, such as reproduction, fighting, food and predators which, as selective factors, dominate all others in their effect, there is little place in the biology of most animals for the kind of subtleties of communication which human language permits. Nor must we forget that communication is a social activity which often runs counter to the trend towards competition which characterizes most animal communities. A very elaborate social organization is required before the survival of an individual's genotype becomes so dependent upon survival of the group that natural selection will encourage individual sacrifice for the sake of the community. In most cases we may expect this to occur only within the family group which is one of the reasons for the strong emphasis on individual information in communication signals. With a more elaborate social organization and division of labor among its members, the immediate pressures upon individual survival are alleviated, and the stage is set for the exploitation of the more subtle gains arising from further elaboration of the systems of communication. The most elaborate communication system known in the animal kingdom occurs in the honey bee, whose social organization particularly from a genetic point of view begins to approach the ideal conditions we have postulated above.

RELATIONSHIPS BETWEEN INFORMATION CONTENT AND SIGNAL STRUCTURE

Human language is usually regarded as consisting of arbitrary symbols, bearing no direct physical relationship to the information which they carry. The communication system is thus based upon a convention. Some zoologists have asserted that the communication signals of animals are arbitrary in the same sense (Lorenz, 1935, 1951) and many of them seem to satisfy the criteria. However, in some cases physical structure is intimately related to the corresponding designata.

The conveyance of locating (i.e. identifying) information by sound signals is directly related to physical structure, since this controls the ease with which the sound source can be located. Vertebrate animals, for example, rely primarily upon differences of intensity, phase and time of arrival of sound at the two ears. The easiest sounds to locate are those providing all of these clues, the ideal being something like a repetitive click. This type of sound, used by many species of birds when they are mobbing an owl (cf. page 308), is a readily located call serving to attract the attention

to the position of the owl. Conversely the calls given when a hawk flies overhead have a different structure which minimizes the clues available for location, making the source of the sound difficult to determine. Insects with different types of receptors which respond, not to pressure changes but to the actual displacement of molecules of the medium, are able to determine the direction of sound directly by reference to the vectorial properties of sound, so that their signals are not affected by the problems of location which confront vertebrates (Marler, 1959).

We have noted that appraisive information is sometimes conveyed by the frequency or length of time for which a signal is repeated. The honey bee dances longer for rich food sources than for poor ones, which implies an iconic relationship.

With the sub-categories of designative information we are on surer ground. Consider for example some of these signals in the light of the most commonly accepted alternative to "arbitrary" which is "iconic". A degree of direct physical correspondence between the signal and its referent is implied (see Cherry, 1957; Hockett, 1961) as with a picture for example. The portion of the honey bee's waggle dance that communicates the direction of the food source conveys this environmental information in an iconic manner by transposing directly from the direction with respect to the sun, to direction with respect to the vertical (Hockett, 1961).

Some other signals conveying environmental information appear to be non-iconic. Thus the various alarm calls of birds bear no relationship to the dangers which are their designata. However their physical structure is by no means arbitrary, in relationship to the locating information which they may convey. Thus the adjectives arbitrary and iconic cannot be applied to a signal as a whole, only to the relationship between signal structure and particular items of information which they communicate.

Signals conveying motivational information may be iconic or arbitrary. Most sound signals probably come into the latter category, although sounds used by many birds and mammals in fighting, having a grating, growling or rattling quality may be related in iconic fashion to the snapping of beaks or teeth which occurs in actual combat. Visual signals which are known to have originated as what zoologists call "intention movements" (Tinbergen, 1940, 1952; Daanje, 1950), which Darwin (1872) recognized as "serviceable associated habits" are more obvious illustrations. For example, many aggressive displays undoubtedly originated through emphasis of the actual physical preparations for attack—baring of the teeth, tensing of the muscles, and so on. More than one type of motivational information may be iconically represented in the same signal, conveying information about the existence of two or more types of motivation in the signaller at the same time. Many insights into the evolution of visual signals have arisen

from the Tinbergen's discovery of this phenomenon of multiple motivation in communicatory behavior.

In the same way signals conveying *negative* motivational information are not arbitrary, since their physical structure is related in an inverse manner to the structure of other signals. These constitute a special class of iconic signals. Finally the need to communicate subtle changes in motivation has repercussions on signal structure, encouraging the use of signals which vary in a continuous manner instead of being discrete, in the appropriate circumstances.

Sexual information may equally well be arbitrary or iconic. The red breast of the reproductive male stickleback, which functions as a signal (Tinbergen, 1951) is arbitrary, whereas the swollen belly of a gravid female, also a signal, is iconic. Individual information also may be arbitrary or iconic. Arbitrariness becomes prominent with species-specific signals. It is no accident that Lorenz's (1951) emphasis on arbitrariness was largely derived from intensive study of the plumage and courtship behavior of ducks and other birds, as they play a role in reproductive isolation, all with a strong emphasis upon species-specific information.

The requirement here is that the signal should be readily distinguished from those of other species likely to be transmitted at the same time and place. The way in which they differ is arbitrary, as long as it is readily perceptible to members of the species. The evidence suggests that this has resulted in specific divergence in a wide variety of animal communication signals which function in reproductive isolation of the species. Even here the signals are not entirely arbitrary, since they are excluded from overlap with the signals of other species.

It will be clear from the above discussion that the classification of signals as either iconic or arbitrary is unsatisfactory. A signal may fail to be entirely arbitrary in several ways, which do not all conform closely to the usual definition of iconic. The structure can, however, be related in different ways to the different types of information being conveyed, be it locating, appraisive, species-specific, environmental, motivational and so on. It may be an aid to further progress if we treat signal structure from this point of view, instead of placing all non-arbitrary signals in the iconic category.

Conclusions on the Evolution of Signals

A detailed review of the evolution of the communication systems of animals is beyond the scope of this paper. We would need to present comparative data, on a much larger scale, and much of the evidence has been reviewed in recent papers (e.g. Tinbergen, 1952, 1959; Morris, 1956, 1957; Marler, 1959) together with discussion of the special problems which arise

with the different sensory modes. We may note that evolution from iconic to arbitrary signals is probably quite a common occurrence, as part of the process known as ritualization (Tinbergen, 1952; Blest, in press). The ontogenetic basis of sound signal systems has been considered in several recent papers (Sauer, 1954; Thorpe, 1958; Messmer & Messmer, 1956; Thielcke-Poltz & Thielcke, 1960; Lanyon, 1957) establishing that while the majority of signals are genetically controlled, some are passed on by the learning of traditions. In contrast we know almost nothing about the ontogenetic basis of responsiveness to signals. Learning probably plays an important role here, even in lower animals. All of these issues need to be considered in a complete analysis of the evolution of the communication systems of animals.

The aim of this essay is more restricted. It seeks only to demonstrate that by using the response evoked by signals as an index, we can derive a picture of the kind of information conveyed. An attempt is made to classify some of the types of information involved, and to show that the effects of natural selection upon the evolution of signals may be clarified by such an approach. The categories suggested are neither final nor exhaustive. The existing knowledge about animal communication is so scanty that we have little to use as a basis. Nevertheless we may make more rapid progress if we approach animal communication systems as a whole instead of treating each aspect as a separate issue. The problems occupy a unique position in the study of the evolution of behavior. It is a challenge for us to try to solve them, even at the most elementary level.

This paper results in part from a research program supported by a grant from the National Science Foundation. It has benefited greatly from the criticisms of Drs. F. A. Beach, C. Cherry, C. F. Hockett and O. P. Pearson.

REFERENCES

Beach, F. A. (1942). *J. comp. Psychol.* **33,** 163.
Beach, F. A. (1947). *Physiol. Rev.* **27,** 240.
Beach, F. A. (1959). *Amer. Psychol.* **15,** 1.
Blest, A. D. In "Modern Problems of the Behaviour of Man and Animals". Cambridge University Press, London. In press.
Cherry, C. (1957). "On Human Communication". Wiley, New York.
Daanje, A. (1950). *Behaviour,* **3,** 48.
Darwin, C. (1872). "The Expression of the Emotions in Man and Animals". Murray, London.
Frings, H., Frings, M., Cox, B. & Peissner, L. (1955). *Wilson Bull.* **67,** 155.
Gallie, W. B. (1952). "Pierce and Pragmatism". Pelican Books, Harmondsworth, England.
Haldane, J. B. S. (1953). *Diogenes,* **4,** 3.
Haldane, J. B. S. (1956). *Science Progress,* **175,** 385.
Haldane, J. B. S. & Spurway, H. (1954). *Insectes Sociaux,* **1,** 247.
Hebb, D. O. (1946). *Psychol. Bull.* **53,** 88.

HEBB, D. O. & THOMPSON, W. R. (1954). In "Handbook of Social Psychology", Vol. 1. (Lindzey, G., Ed.) Addison-Wesley, New York.
HINDE, R. A. (1955–6). *Ibis*, **97**, 706; **98**, 1.
HOCKETT, C. F. (1961). In "Animal Sounds and Communication". (Lanyon, W. E. & Tavolga, W. N., Eds.) A.I.B.S. Symposium Proceedings. A.I.B.S. Washington.
KOHTS, N. (1935). *Sci. Mem. Mus. Darwin., Moscow*, **3**, 1. Russian with English summary.
LANYON, W. E. (1957). *Publ. Nuttall Ornithol. Cl.* **1**, 1.
LEHRMAN, D. S. (1953). *Quart. Rev. Biol.* **28**, 337.
LEYHAUSEN, P. (1956). *Handbuch Zool., Berlin, Bd. VIII*, **10**(21), 1.
LINDAUER, M. (1957). *Nature, Lond.* **179**, 63.
LORENZ, K. Z. (1935). *J. Ornithol.* **83**, 137, 289.
LORENZ, K. Z. (1951). *Symp. Soc. exp. Biol.* **4**, 221.
MARLER, P. (1956). *Ibis*, **98**, 231.
MARLER, P. (1959). In "Darwin's Biological Work". (Bell, P. R., Ed.) Cambridge University Press, London.
MARLER, P. In "Vertebrate Speciation". (Blair, W. F., Ed.) University of Texas Press. In press.
MESSMER, E. & MESSMER, I. (1956). *Z. Tierpsychol.* **13**, 341.
MORRIS, C. W. (1946). "Signs, Language and Behavior". Prentice Hall, New York.
MORRIS, D. (1955). *Behaviour*, **8**, 46.
MORRIS, D. (1956). "The function and causation of courtship ceremonies". Fondation Singer Polignac Colloque Internationale sur L'Instinct, 1954.
MORRIS, D. (1957). *Behaviour*, **11**, 1.
MOYNIHAN, M. (1955). *Behaviour Suppl.* **4**, 1.
NICE, M. M. (1943). *Trans. Linn. Soc. N.Y.* **4**, 1.
OGDEN, C. K. & RICHARDS, I. A. (1923). "The Meaning of Meaning". Routledge and Kegan Paul, London.
POTTER, R. K., KOPP, G. A. & GREEN, H. C. (1947). "Visible Speech". van Nostrand, New York.
REVESZ, G. (1956). "The Origins and Prehistory of Language". Longmans, Green & Co., London.
SAUER, F. (1954). *Z. Tierpsychol.* **11**, 10.
SHANNON, C. E. & WEAVER, W. (1949). "The Mathematical Theory of Communication". University of Illinois Press, Urbana.
SIBLEY, C. R. (1957). *Condor*, **59**, 166.
SMITH, S. S. & HOSKING, E. (1955)."Birds Fighting". Faber & Faber, London.
STEVENS, S. S. (1950). *J. acoust. Soc. Amer.* **22**, 689.
THIELCKE-POLTZ, H. & THIELCKE, G. (1960). *Z. Tierpsychol.* **17**, 211.
THORNDIKE, E. L. (1911). "Animal Intelligence". Macmillan, New York.
THORPE, W. H. (1956). "Learning and Instinct in Animals". Methuen, London.
THORPE, W. H. (1958). *Ibis*, **100**, 535.
TINBERGEN, N. (1940). *Z. Tierpsychol.* **4**, 1.
TINBERGEN, N. (1951). "The Study of Instinct". Clarendon Press, Oxford.
TINBERGEN, N. (1952). *Quart. Rev. Biol.* **27**, 1.
TINBERGEN, H. (1959). *Behaviour*, **15**, 1.
TINBERGEN, N. & MOYNIHAN, M. (1952). *Brit. Birds*, **45**, 19.
TOLMAN, E. C. (1932). "Purposive Behavior in Animals and Man". Appleton-Century-Crofts, New York.
TRAGER, G. L. (1958). *Stud. Linguistics*, **13**, 1.
VON FRISCH, K. (1954). "The Dancing Bees". Methuen, London.

J. Theoret. Biol. (1961) 1, 318–323

Experimental Examination of the Hypothesis of Coupling between the Sodium and Potassium Pump in Erythrocytes

W. Rummel, K. Pfleger and E. Seifen

The Pharmacological Institute of the University of the Saarland, Homburg/Saar, Germany

(*Received* 30 *December* 1960)

The effect of iodoacetate and arsenate upon movement of sodium and potassium in rabbit blood has been investigated. The experiments of Scott & Hayward in *Ulva lactuca* were repeated in erythrocytes: arsenate prevents the K loss from red cells induced by iodoacetate without an effect upon Na increase in the cells (as measured by flame-photometer). However, separate analysis of influx and efflux by radioisotopes provides no evidence that arsenate activates K influx, when reduced by iodoacetate. Arsenate inhibits the K loss under iodoacetate by decreasing the K efflux. Influx of sodium is also inhibited by arsenate, but not as strongly. Thus the movement of both ions in the direction of decreasing electrochemical potential gradient is inhibited by arsenate. The greater action on K movement is compatible with the assumption that the passive movement of ions depends upon the radius of the non-hydrated ions not of the hydrated ions. It cannot be concluded from the experiments presented that the transport mechanisms for Na and K are independent of one another.

Introduction

At the present time it is usually assumed that the uphill movements of sodium and potassium ions in erythrocyte membranes are strongly coupled to one another (Post & Jolly, 1957; Glynn, 1959). In earlier discussions of this question evidence was often quoted that showed that, under certain conditions, Na and K can move independently (Davson & Reiner, 1942; Solomon, 1953). Harris (1941) and later Maizels (1948) demonstrated that erythrocytes, preincubated at 7°C, when suspended in a solution without glucose, can still accumulate potassium, while the sodium concentration does not decline but increases from the very beginning. Wilbrandt (1940) and Dunker & Passow (1950) were able to find conditions in which the opposite result is seen; the cells lose potassium while the sodium gradient is maintained. Scott & Hayward (1953, 1955) have recently published observations in algae (*Ulva lactuca*) which also seem to indicate that the sodium and potassium pumping mechanisms can work independently. It was of obvious interest to see if the results could be reproduced under similar conditions in erythrocytes.

Methods

Blood was obtained from rabbits by puncture of the central artery of the ear, and was defibrinated by gentle stirring with wooden sticks. Samples of 1 ml of blood each were incubated for 60 minutes at 37°C in 15 ml centrifuge tubes. Iodoacetate and Na_2HAsO_4 dissolved in 0·9% NaCl solution was added in a volume of 0·1 ml; pH values at the end of the experiments were between 7·5 and 7·6. The solutions were shaken at intervals.

NET FLUX

The external potassium concentration was determined in the supernatant by flame-photometry after addition of 10 ml ice-cold isotonic $CaCl_2$ solution and centrifugation. Before determination of the sodium concentration the washing procedure was repeated three times. The washed erythrocytes were hemolysed by addition of 5 ml distilled water and sodium was determined in a flame-photometer ("Eppendorf", Fa.Netheler & Hinz, Hamburg). The readings of the instrument for sodium were multiplied by a factor of 1·2 since control experiments had shown that the indication with solutions containing hemolysed erythrocytes was 20% lower than in clear solutions. In each run three samples were taken before incubation; mean values of the K concentration in the serum and Na concentration in the erythrocytes were determined and subtracted from the values obtained at the end of the experiment, to give net flux.

SODIUM AND POTASSIUM INFLUX

To determine the influx of sodium and potassium the procedure described above was followed, but to each flask a volume of 0·1 ml of 0·9% NaCl solution was added containing carrier-free ^{22}Na (39,500 counts/min in our counting arrangement) and ^{42}K (41 μg K ; 120,000 counts/min). The erythrocytes were washed four times at the end of the experiment in 50 ml ice-cold NaCl solution and hemolysed with 5 ml distilled water. Two milliliters of each hemolysed sample were dried on plates (3 cm diameter), and the activity of the β-radiation of each sample determined (methane-flow counting tube with end window). Na activity was eliminated by means of an aluminum foil of 130 mg/cm^2. Later, after the activity of ^{42}K had dropped to insignificant values Na activity was determined under the same geometrical conditions but without aluminum foil. The absolute amounts of Na and K movement were calculated from the activity counted, taking into consideration the specific activity and the decay of activity.

SODIUM AND POTASSIUM EFFLUX

To determine efflux, 10 ml of blood were incubated with 1 ml of the ^{22}Na, ^{42}K solution for 90 minutes at 37°C. After four washings with ice-cold NaCl solution the erythrocytes were suspended in 6 ml of serum from the same animal. The inhibitors were dissolved in saline solution and were added in a volume of 0·1 ml to the recomposed blood. After 60 minutes at 37°C 2 ml saline were added, the suspension was centrifuged, the activity of ^{22}Na and ^{42}K in the supernatant was counted and flux was calculated as described.

The net flux values were subjected to the t-test (see Table 2). The values for influx and efflux were paired and the differences used for t-test.

Results and Discussion

The distribution of sodium and potassium in rabbit blood is influenced by iodoacetate and by arsenate, and by a mixture of both inhibitors, in principle in the same way as described for *Ulva lactuca* (Table 1). In the

TABLE 1

Net flux of sodium and potassium in rabbit erythrocytes under the influence of metabolic inhibitors (expressed as meg/l erythrocytes; incubation period 60 min, pH 7·5, 37°C)

	Controls	Iodoacetate 5×10^{-4} M	Arsenate 5×10^{-3} M	Iodoacetate 5×10^{-4} M + arsenate 5×10^{-3} M
Sodium increase ($n = 10$)	0·07 ± 0·16	1·39 ± 0·45	0·96 ± 0·34	1·65 ± 0·57
Potassium loss ($n = 10$)	1·06 ± 0·90	6·25 ± 1·04	0·68 ± 0·55	1·92 ± 0·95

P = 0·01 (except for the relation controls←→arsenate)

presence of iodoacetate the erythrocytes lose potassium, as does *Ulva lactuca* (Scott & Hayward, 1953, 1955). The potassium loss is three to four times higher than the amount of sodium taken up simultaneously. The only difference between the present results and those with *Ulva lactuca* is a small gain in sodium content seen under arsenate in erythrocytes. This effect, however, is considerably smaller than with iodoacetate.

Crucial for the discussion of the hypothesis of coupling between sodium and potassium transport in erythrocytes is the question whether only the increased K loss by iodoacetate can be reduced by arsenate or whether the uptake of sodium is also reduced. The findings with erythrocytes are the same as those obtained with *Ulva*. The K loss under iodoacetate is reduced

by the simultaneous presence of arsenate, and the increased sodium uptake remains unchanged. Scott & Hayward concluded from their results that the mechanisms responsible for accumulation of potassium and secretion of sodium respectively are independent of one another (see also the review of Glynn, 1959). It is tempting to use our results as argument for the hypothesis that also in erythrocytes the sodium and potassium pump are not strongly coupled. Following Scott & Hayward one would have to assume that in the presence of arsenate the iodoacetate block of glycolysis can be bypassed by "arsenolysis" (Warburg & Christian, 1939) and that the metabolic steps concerned are coupled only to the potassium transport system. In other words, under these conditions only the trans-

TABLE 2

Influx and efflux of potassium and sodium in rabbit erythrocytes under the influence of metabolic inhibitors (expressed as meg/l erythrocytes; incubation period 60 min, pH 7·5, 37°C)

	Controls	Iodoacetate 5×10^{-4} M	Arsenate 5×10^{-3} M	Iodoacetate 5×10^{-4} M +arsenate 5×10^{-3} M
Sodium influx	6·39	7·15	6·72	6·78
Sodium efflux	6·07	5·55	5·79	5·58
Potassium influx	2·10	1·17	1·52	1·02
Potassium efflux	2·64	3·90	2·61	3·11

Test of significance:

Na influx:	K↔J	: $\bar{d}$ =	+0·76,	s = 0·455,	p = 0·01
	K↔As	: $\bar{d}$ =	−0·34,	s = 0·670,	p = 0·1
n = 9	K↔J+As	: $\bar{d}$ =	+0·39,	s = 0·630,	p = 0·05
	J↔J+As	: $\bar{d}$ =	−0·35,	s = 0·428,	p = 0·05
Na efflux:	K↔J	: $\bar{d}$ =	−0·51,	s = 0·337,	p = 0·01
	K↔As	: $\bar{d}$ =	−0·28,	s = 0·255,	p = 0·02
n = 9	K↔J+As	: $\bar{d}$ =	−0·49,	s = 0·315,	p = 0·01
	J↔J+As	: $\bar{d}$ =	0		
K influx:	K↔J	: $\bar{d}$ =	−0·93,	s = 0·117,	p = 0·01
	K↔As	: $\bar{d}$ =	0·58,	s = 0·018,	p = 0·01
n = 4	K↔J+As	: $\bar{d}$ =	−1·08,	s = 0·068,	p = 0·01
	J↔J+As	: $\bar{d}$ =	−0·15,	s = 0·054,	p = 0·01
K efflux:	K↔J	: $\bar{d}$ =	+1·29,	s = 0·287,	p = 0·01
	K↔As	: $\bar{d}$ =	0		
n = 4	K↔J+As	: $\bar{d}$ =	+0·52,	s = 0·225,	p = 0·01
	J↔J+As	: $\bar{d}$ =	−0·79,	s = 0·250,	p = 0·01

K = controls
J = iodoacetate
As = arsenate

$\bar{d}$ = mean of the difference
s = standard deviation of d

port mechanism for potassium would be reactivated, while the sodium pump would remain inhibited.

The theoretical consequences of such a statement would be far-reaching. However, until now the statement is based only upon results obtained with flame-photometric technique, that is only upon determinations of net fluxes. To strengthen these conclusions it was necessary to test the effect of both inhibitors upon influx and efflux with the help of radioisotopes.

The results of such experiments are summarized in Table 2. The experiments showed that, with regard to the movement of Na and K *against* the concentration gradient, iodoacetate inhibits K influx, as compared to controls, by almost 50%, while the efflux of sodium is decreased by only about 10%. Simultaneous presence of arsenate, however, does not reverse either the inhibition of the K uphill transport or that of Na. Rather, the inhibition of the K uphill transport is increased, although only slightly, by addition of arsenate.

The movement of Na and K *with* the concentration gradient is accelerated under iodoacetate. This effect of iodoacetate is inhibited by arsenate. The increased K efflux is reduced by 20%, the Na influx by 5%. (For technical reasons the experiments concerning efflux and influx were not carried out under exactly indentical conditions. Therefore, the results can be compared only within each group of experiments. This comparison suffices, however, to answer the question posed.)

The question initially stated regarding the cause of the action of arsenate upon K loss under iodoacetate can now be answered. The inhibition of efflux is the only factor responsible for the observed net change in K distribution. The inhibition of K influx caused by iodoacetate is *not* reversed by arsenate. There is no reason to assume that this explanation is not valid for the findings with *Ulva lactuca* as well.

It seems noteworthy that arsenate affects mainly the action of iodoacetate upon ionic movements *with* the concentration gradient. Apparently arsenate reduces the increase in membrane-permeability caused by iodoacetate. The greater effect upon K movement can be understood if one assumes—as does Mullins (1959)—that the effective ionic radius for the passage of ions through the pores of the membrane is that of the non-hydrated ion, not that of hydrated ion. Such an assumption would also provide an explanation for the finding by Harris (1954) that the passive permeability of erythrocytes for Na is twice that for K. Similar results have also been obtained for the permeation of Na through the intestinal wall (Rummel & Stupp, 1960).

Although these experiments have failed to provide evidence that the Na and K pump in erythrocytes work independently of one another, neither do they furnish any proof that a strong coupling does exist. Exchange one

for one, which is presumed by Glynn (1959) and Post & Jolly (1957) for human erythrocytes, certainly does not occur in rabbit red cells (Pfleger, 1960; Seifen, 1960). The fact that the K loss of erythrocytes in iodoacetate is three to four times as high as the Na uptake equally does not support the thesis of one-to-one exchange. Movement of potassium, *with* the concentration gradient as well as *against* it, is inhibited by iodoacetate more strongly than is movement of sodium.

Thus the experiments have not provided an answer to the question of how, if at all, uphill transports of sodium and potassium in the erythrocyte membrane are linked. They may, however, aid the development of a final theory concerning the coordination of K and Na transport by eliminating one hypothetical mechanism.

We wish to thank Dr. Werner Flacke for translation and discussion.

REFERENCES

Davson, H. & Reiner, J. M. (1942). *J. cell. comp. Physiol.* **20,** 325.
Dunker, E. & Passow, H. (1950). *Pflüg. Arch. ges. Physiol.* **252,** 542.
Glynn, M. (1959). *Int. Rev. Cytol.* **8,** 449
Harris, E. J. (1954). In "Symposia of the Society for Experimental Biology", No. VIII. Cambridge University Press, London.
Harris, J. E. (1941). *J. biol. Chem.* **141,** 579.
Maizels, M. (1948). *J. Physiol.* **108,** 247.
Mullins, L. J. (1959). *J. gen. Physiol.* **42,** 1013.
Pfleger, K. (1960). "Die Beziehungen zwischen Ionenaustausch und Stoffwechsel". Vortrag: *Tag. Ges. physiol. Chem. Zürich.*
Post, R. L. & Jolly, P. C. (1957). *Biochim. biophys. Acta*, **25,** 118.
Rummel, W. & Stupp, H. F. (1960). *Arch. exp. Path. Pharmak.* **240,** 79.
Scott, G. T. & Hayward, H. R. (1953). *J. gen. Physiol.* **36,** 659.
Scott, G. T. & Hayward, H. R. (1955). In "Electrolytes in biological systems". (A. M. Shanes, Ed.). American Physiological Society, Washington.
Seifen, E. (1960). "Die pH-Abhängigkeit des Ionendurchtritts durch die Erythrozytenmembran". Vortrag: *Tag. Ges. physiol. Chem. Zürich.*
Solomon, M. K. (1953). *J. gen. Physiol.* **36,** 57.
Warburg, O. & Christian, W. (1939). *Biochem. Z.* **303,** 40.
Wilbrandt, W. (1940). *Pflüg. Arch. ges. Physiol.* **243,** 519.

J. Theoret. Biol. (1961) 1, 324–327

Division of Labor in a Honey Bee Colony—a Markov Process?

Adrian M. Wenner

Department of Biological Sciences, University of California, Santa Barbara University, California, U.S.A.

(*Received* 20 *January* 1961)

Introduction

Individual honey bees within a hive perform activities that contribute to the well-being of the colony. The independent performance of these activities generally leads to the maintenance and growth of the colony and gives the impression of order and directiveness. Continued observation of one individual for a several-hour period, however, shows that a worker may change activities frequently and without apparent direction. Cell construction, for instance, may be followed by feeding larvae, by accepting nectar from foragers, or by any one of many possibilities.

Not only does one activity not appear to be a precursor to another activity, but activities by one individual do not always appear to influence activities of another individual. Immediately after one worker has molded wax into the wall of a cell, for example, another worker may remove the same wax and apply it to another cell. In spite of this changing from activity to activity and frequent undoing of that which is already done, the collective activity of individuals results in an ultimate satisfaction of colony requirements. Butler (1949) summarized this dilemma as follows: "It is clear that worker honeybees can shift from one job to another as the requirements of their colony demand, but we still know little concerning the method by which each worker discovers the particular job that she is called upon to perform at any given time in order to help to maintain the economy of her colony in the most efficient manner."

Contributions to the Problem

The continued interest shown in this problem is reflected in the review by Ribbands (1953, Chapter 37). He discusses the widely quoted report of Gerstung, who "... postulated a detailed scheme for allocation of duties according to age". Subsequently, according to Ribbands, Rösch supported the conclusion of Gerstung after watching marked bees which

had been introduced into nuclei (small hives). Perepelova, Ribbands, and Lindauer each repeated the study with a similar technique (this is discussed in detail by Ribbands). Whereas the results of Perepelova agreed with those of Rösch, the results of Ribbands and Lindauer indicated a much greater plasticity. This variability has been repeatedly demonstrated by the success of those who have attempted to start nuclei with newly emerged bees—bees that have had no contact as adults with adults of the previous generation. Although Ribbands does not stress the point, this plasticity is also evident from the observation that older, over-wintered bees engage in the entire spectrum of activity during their spring build-up. More specifically, Meyer (1956) later found that a newly settled swarm contained wax-producing bees of all ages. Ribbands summarized his review by stating that bees have "... a marked preference for one or other of the various tasks, and that this *preference* (Ribband's italics) changes with age although all the tasks can still be carried out". He adds that "... the allocation of duties is arranged to meet the needs of the community, whatever they may be".

Lindauer (1953) also used the technique of observing one bee during its entire life of 24 days and a second for its first 8 days of life. In addition, the activities of a third were anticipated and an attempt was made to eliminate any work for this bee. Observations on this third individual were continued without interruption for 8 days. These observations indicated that bees constantly shift from activity to activity without obvious cues from other bees or from the environment (with the remarkable exception of those bees which respond to the signal given by dancing bees) and gradually emphasize different activities as they age. Lindauer says, "The life histories of the three bees mentioned show clearly that age does exert a strong influence on the division of labour for the individual bee, but that the determining factor seems to be the needs of the colony at the time." He states, "The chief aim of future investigations must be to try to find out how the bees know at this or that moment that a certain type of activity is needed." He later modifies this statement somewhat by concluding, "The last question is therefore no longer 'How does the individual bee *know* what to do next in order to supply the needs of the colony as a whole?' but 'How does it come about that all the firmly fixed instinctive actions and reactions—which in the thousands of separate bees respond to the diverse stimuli of the environment—all fit together into one harmonious whole?"

The findings of Sakagami (1953a, b) and Istomina-Tzvetkova (1953, 1957) are in agreement with those of Lindauer and Ribbands: bees do not follow a regular sequence in their activities within the hive, and activities within the hive are somehow determined by the needs of the colony. Other studies have concentrated on some particular aspect of the problem. Free

(1959) reviews findings from work done on the importance of food transfer, and Butler (1959) reviews studies on effects of "Queen substance" on the colony. But the question of division of labor within a colony still remains unsolved.

Even in such a familiar phenomenon as the circle of bees surrounding the queen, no one yet knows what determines which bees are to be present or what determines the action of the individual bees when they are in contact with the queen. Allen (1960), for instance, in a study of queen attendants, concluded, "the behaviour of individual bees varied on different visits, but no regular sequence could be found".

The Markov Process Possibility

Approaches to the problem of organization with a honey bee colony have been deterministic and have followed the chain-reflex theory of behavior. If reflex chains governed a community so complex as that of honey bees, however, the pattern of sequences would probably be grotesque in its complexity. The fact that virtually no behavioral sequences have been found suggests that few exist and that some alternative to the chain-reflex theory might better explain the efficient functioning of this community. Such an alternative must account for the following phenomena within the hive:

1. Not all bees are active at any one time.
2. Bees may work for only a few minutes and/or do not finish one activity before changing to another.
3. One bee may undo the work of another.
4. Not all bees react to obvious stimuli such as dancing bees or hive intruders.
5. The transition from activity to activity occurs in the apparent absence of cues.
6. As bees age they tend to emphasize different activities.
7. Total emphasis within the hive changes as the internal environment of the hive changes.
8. Under conditions of stress bees can engage in any occupation, regardless of age; and therefore, a nucleus of newly emerged bees can succeed.

A cybernetic approach, as suggested for the behavior of populations by Ashby (1956, Chapter 9), can be made to this problem. If the system is considered to be controlled by a Markov process, the 8 points listed are explainable by the following three assumptions (inherent in a stochastic process):

1. The transition from one activity to another takes place at random (points 2, 3, 4, and 5).

2. The probabilities of transitions to different activities differs, dependent on the nature of the two activities involved (points 6 and 7).
3. The probability of the same transition changes with the age of the individual or under environmental change (points 6, 7, and 8).

At the same time, the failure of students of the problem (to find behavioral sequences) becomes understandable.

A honey bee colony is appropriate for the investigation of the existence of such a process. The colony size can be regulated, environmental conditions can be fairly uniformly controlled, and individual bees can be marked and watched. A Markov matrix of probabilities can be constructed on the basis of observation; and environmental factors can then be altered one at a time and the effect on the matrix can be studied. In addition, any behavioral sequences that are present will become obvious from the matrix, in that the particular transition will have a probability of 1. Since published data is inadequate for this type of analysis, an experimental program designed to test the stochastic hypothesis of hive functioning is now in progress.

I am grateful to Eduardo Orias for providing the critical stimulus which led to the development of this paper. I am also grateful to the student and faculty members of our department who read this paper and who generously contributed their time and consideration to the argument.

REFERENCES

Allen, M. D. (1960). *Anim. Behav.* **8,** 201.
Ashby, W. R. (1956). "An Introduction to Cybernetics". Chapman and Hall, London.
Butler, C. G. (1949). "The Honeybee (An Introduction to her Sense-physiology and Behaviour)". Clarendon Press, Oxford.
Butler, C. G. (1959). *Bee World,* **40,** 269.
Free, J. B. (1959). *Bee World,* **40,** 193.
Istomina-Tzvetkova, K. P. (1953). *Pchelovod,* **9,** 15. In Russian. Trans. (1954) M. Simpson in "Some Recent Russian Researches on Bees and Beekeeping". Bee Research Association, London.
Istomina-Tzvetkova, K. P. (1957). *Zool. Zh.* **36,** 1359. In Russian. Abstract No. 360/59 in *Bee World,* **40**
Lindauer, M. (1953). *Bee World,* **34,** 63, 85.
Meyer, W. (1956). *Insectes Sociaux,* **3,** 303.
Ribbands, C. R. (1953). "The Behaviour and Social Life of Honeybees". Bee Research Association, London.
Sakagami, S. F. (1953a). *J. Fac. Sci. Hokkaido Univ. (Zool.)* **11,** 343.
Sakagami, S. F. (1953b). *Jap. J. Zool.* **11,** 117.

J. Theoret. Biol. (1961) **1**, 328–341

Physical Aspects of Protein and DNA Synthesis

Ernest Pollard

College of Chemistry and Physics,
The Pennsylvania State University, U.S.A.

(*Received* 23 *January* 1961)

The idea that the synthetic processes of a bacterial cell take place by random collision and specific selection has been examined in terms of Brownian movement theory. The observed rates of protein, deoxyribonucleic acid and ribosome synthesis are used to calculate the necessary concentrations of the metabolites used in the synthetic process. If the collision process must involve the presentation of the active part of the colliding molecule at first impact, then the observed rates require excessive concentrations for soluble ribonucleic acid. The concentration required for nucleotide triphosphates in the synthesis of deoxyribonucleic acid is such that many points of synthesis are called for.

If, instead, a very temporary complex can form, of the nature expected from the operation of London–Van der Waals forces, the free rotation of the colliding molecule provides a variety of aspect changes, and the collision process is much more efficient. If this process takes place, the transfer of energy from rotation to vibration becomes important.

The precision of assembly of deoxyribonucleic acid is examined in terms of the idea that precision in timing is important. It is shown that the cell could not function in that way if any large molecule had to move during the assembly.

It is suggested that enzymes having high activation energies must be present in abnormally high concentrations.

Introduction

In recent years two of the most important synthetic processes in the cell have begun to receive descriptions which are relatively detailed. The process of *protein synthesis* is supposed to occur on ribosomes, particles of ribonucleoprotein. Recent references include Dintzis, Borsook & Vinograd (1959) and McQuillen, Roberts & Britten (1960). The specific ordering of the amino acids is supposed to be conferred by a base-pairing between the ribonucleic acid (RNA) of the particles and soluble RNA combined specifically with each amino acid (Crick, 1958). The amino acids, prior to combination, are activated by combination with adenosine triphosphate (ATP), using a specific enzyme for each amino acid (see Hoagland, Keller & Zamecnik, 1956; Zamecnik, Stephenson & Hecht,

1958). The process of deoxyribonucleic acid (DNA) *synthesis* has been shown by Kornberg and his group (Bessman, Lehman, Simms, & Kornberg, 1958) to be produced from the triphosphates of the four constituent bases, by a "polymerase", using single strand DNA as a primer.

Both these processes are very accurate and represent, in a certain sense, the peak of precision performance by the cell. Regarded as an assembly process, the accuracy achieved is remarkable, because the supply line is supposed to be random bombardment—the Brownian motion of the metabolites; and the forces acting are supposed to be those of normal chemical bonding, plus Van der Waals forces, and not specifically operating forces of long range. One of the key questions posed by the resolute advance of biochemistry into the inner functioning of the cell is whether the processes discovered by biochemistry can actually give a description of the working of the cell. It is not possible to answer such a question here, but it is feasible to make a limited physical analysis of the process of *random collision and specific selection* to see whether any physical inconsistencies exist. It will be seen that, at present, no such inconsistencies can be developed, but nevertheless some requirements are imposed on the process, and the importance of one form of energy transfer shows up.

The examination of one feature of synthesis, collision and specific capture, will be applied to two stages of protein synthesis; to DNA synthesis; and to the supposed combination of protein and RNA to form ribosomes. The precision of formation of DNA will be examined assuming the accuracy to depend on speed, and the process of change from rotation to vibration, which emerges as an important feature, will be briefly considered.

Special Terms and Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
Ribosome	Particulate combination of RNA and protein and a probable site for protein synthesis
London–Van der Waals forces	Nonspecific attractive forces occurring between all types of molecule. This type of force is cumulatively additive, in contrast to a valence force or a hydrogen bond
Activation energy	An energy that must be overcome before some reaction occurs
ATP	Adenosine triphosphate
Å	Ångström unit, 10^{-8} cm
Bit	Unit of information. One yes or no choice

R	Molecular radius of molecule undergoing collision
C	Concentration in molecules per cc
C_0	Concentration far from a reacting site
D	Diffusion coefficient in cm^2/sec
ϕ	Number of reacting molecules at one reaction site per second
g	Ratio of active area to total area for a colliding molecule
k_1	Rate constant for formation of Enzyme-Substrate complex
k	Boltzmann's constant, $1{\cdot}37 \times 10^{-15}$ erg/°C
T	Absolute temperature
η	Coefficient of viscosity, measured in poises
AA	Amino Acid
AAA	Activated Amino Acid
SRNA	Soluble, or transfer RNA
Template	Postulated ordering region for protein synthesis
Θ	Rotational diffusion coefficient in $(radians)^2/sec$
ΔE	Inherent uncertainty in energy
Δt	Inherent uncertainty in time
h	Planck's constant

Experimentally Observed Features of Cell Operation

The discussion is primarily in terms of the bacterium *Escherichia coli*, but can be extended to both viruses and to larger cells than bacterial cells, if need be.

Under optimum conditions, cells of *E. coli* divide in 20 minutes. In synthetic medium the time is 50 minutes. The cells are 0·85 micron in diameter, variable in length, with 2·5 microns as a representative value. Hanawalt (1958) found a DNA content of $0{\cdot}75 \times 10^{-14}$ gram per cell and an RNA content of $4{\cdot}4 \times 10^{-14}$ gram per cell. The total dry mass per cell is 25×10^{-14} gram. Assuming protein to be 60% of the dry weight, the mass of protein is 15×10^{-14} gram. Assuming the RNA in a ribosome has a molecular weight of 2×10^6, then a number of 10^4 ribosomes per cell is reasonable. The RNA per cell is not a fixed number, but varies with the conditions of growth. Hence 10^4 ribosomes is not an exact number.

There is no clear evidence for a cycle of synthesis in *E. coli*. We therefore propose to use the whole period of division as the time for synthesis. If cycles are subsequently found, the rates will be higher, at maximum speed, than used here.

Taking the average molecular weight of an amino acid as 150 and of a base to be 300, the above figures require that, for a division time of 1,200 seconds, 5×10^5 amino acids and $1{\cdot}2 \times 10^4$ nucleotides be assembled per second. Since not all amino acids are equally represented, we take a representative fastest rate of amino acid assembly to be 100/sec/ribosome.

The viscosity of the material inside a bacterial cell is most important. A recent experiment by Dr. F. L. Gardner of the Yale Biophysics Department, showed that cell "juice" produced by opening cells under pressure is very viscous, with a very interesting dependency on temperature, the minimum value, at 37°C being 600 centipoises. Some of this is undoubtedly due to the presence of partly broken down DNA, which would be organized in the nucleus of the cell. It does not seem unreasonable, however, to assume a viscosity of 1 6 of this measured amount, or 100 centipoises. Whether this figure, or the lower value for the microscopic viscosity should be used is a question. Additional allowance for collisions before reaching the final synthetic region would be needed if the microscopic value were used.

Random Collision and Specific Selection

The process considered is twofold. It comprises a random Brownian motion, which, by chance, causes many metabolites to collide with many parts of the cell, at a rate calculable from the theory of Brownian motion. It also includes the idea that a collision made at the right place, by the right molecule and presenting the right aspect, will result in a capture of the molecule and an inclusion of it in the macromolecule being synthesized. These two processes form the scheme of synthesis we call *random collision and specific selection*. Both processes must now be discussed.

SPECIFIC SELECTION

It is generally accepted, and will not for the moment be questioned, that "specificity" resides in a similar or complementary molecular shape. For combination to result after the contact of such fitting surfaces there must also be forces, which are small for the wrong shape, and decisively large for the right shape. The forces available are, in order of magnitude: the covalent bond; ionic bonding (not too easy to apply to general biological combinations), and hydrogen bonding. All these extend over 1 or 2 Å. There is a fourth interaction: the London–Van der Waals force. This last force differs from the three preceding in that it does not have the feature of saturation. It is due to the effect of a transient dipole on nearby atomic systems, in which dipoles are induced by the transient dipole. All such interactions are *additive*, at least for small distances of interaction, and, as a result, the London force exerts a general, nonspecific force, not dependent upon aspect and of relatively long range. Calculations made by Vold (1954) on the interaction of objects of various shapes indicate that, for distances less than the cube root of their volume, two particles attract each other with energies greater than thermal agitation.

All four of these forces thus have the property that they are of short

range. The London–Van der Waals force can extend to several Ångström units, if the particles involved are themselves large.

To describe specificity we can therefore apply the first and obvious requirement, that there is only one class of molecule that will fit. However, for a collision process we have also to ask whether every collision succeeds. In other terms, does every aspect of approach cause a successful encounter? Must an amino group encounter a carboxyl or a phosphorus atom a hydroxyl on a sugar? We intend to suppose as a hypothesis that correct aspect *is* important and that two relatively small regions on the colliding molecule and on the synthetic surface must approach within the range of the forces we have described in order to complete the process. Estimates of the size of these "relatively small" regions are hard to make. For one approach we follow Quastler (1953) who concluded that biological specificity very often occurs with information content of 7 to 9 bits. Expressing 7 binary choices as a fraction we obtain 1/128. For our purpose we can say that 1/100 of the area of the colliding molecule is specifically related to the synthetic surface. The fraction may be too small if the specificity is not wholly in the molecular surface. However we can use Quastler's figure for this analysis. For another approach we can consider a typical synthetic process—the attachment of a nucleotide triphosphate on to a forming DNA chain. The nucleotide triphosphate, of molecular weight 450, has an area of roughly 300 Å^2. The attachment occurs at a particular phosphorus atom for which we can guess an area of 4 Å^2. The ratio is 1/75, not absurdly different from that obtained from the figure of 7 bits. Clearly "specificity" will be expected to be variable. We propose, as a means of starting, to assume that 1/100 of the surface of a molecule is "active" and that the rest is inert.

RANDOM COLLISION

The first treatment of three-dimensional Brownian movement as applied to this class of problem was that of Smoluchowski (1916), who applied it to the theory of rapid coagulation of colloids. In his treatment the equivalence between diffusion processes involving boundary conditions and Brownian movement was used (see Chandrasekhar, 1943) and a very simple result deduced. The process of collision with a target molecule of radius R was regarded as removing completely the colliding molecule so that, if C is the concentration, in molecules per unit volume, of the colliding molecules, the first boundary condition is that for a radius R, $C = 0$. The second is that, at infinity from the target molecule $C = C_0$. Under these circumstances (see Overbeek, 1952) it is not hard to show that the rate of reaction, ϕ, in molecules per second is

$$\phi = 4\pi DRC_0 \qquad (1)$$

where D is the sum of the diffusion coefficients of the colliding and the target molecules.

Coagulation of colloids is a particularly simple type of process, in that every collision can be taken as effective. Our case differs, in that a majority of collisions will take place with the wrong aspect, and these will not remove the colliding molecule. Something like this was considered by Collins & Kimball (1949), who showed that an efficiency factor can be introduced into equation (1) without changing the form of the equation. The treatment they use probably cannot be applied rigorously to our case, because the efficiency factor introduced is dimensionally a velocity, whereas the concept needed for biosynthesis is the correct approach of two small surfaces. The treatment of Collins & Kimball is, however, of great value in that it shows that the necessary concentration gradient to provide diffusion is not dependent on the removal of the colliding molecule *at every collision*, and therefore it strengthens the validity of the simple approach we propose to use.

In every case that has to be considered, the specific element of synthesizing or "activating" surface is on a relatively large structure, either an enzyme molecule or a ribosome. We propose to consider the radius in the use of equation (1) to be that of the colliding molecule, and, following Strother & Ackerman (1961), to treat the element of surface on the enzyme or ribosome as accessible only from one side, which reduces the factor 4π to 2π. The property of specificity, as discussed earlier, introduces a factor, which we term g, which is the ratio of active area to total area, and which we propose to treat as 1/100.

There are now two methods we propose to use to calculate the actual rate of synthesis. The first assumes that the colliding molecule, in order to make a successful collision, must present the proper aspect, essentially at collision. The second assumes that upon collision the colliding molecule is loosely held, long enough for the rotational Brownian motion of the held molecule to have a real influence on the presentation of the active element of molecular surface to the synthetic site. For the first method, equation (1) is readily modified to read

$$\phi = 2\pi DRgC_0 \tag{2}$$

where ϕ is the number of molecules processed at a single processing site (whether enzyme or ribosome), D is the sum of the diffusion constants of the reacting elements, R is the radius of the colliding molecule, g is the specificity factor and C_0 is the concentration of the colliding molecule in numbers of molecules per unit volume.

For the second method we assume that the process of attachment occurs, as in the rapid coagulation of colloids, at the rate given by equation (1),

which lacks the g factor. The colliding molecule is, however, loosely held and not specifically combined. While held, it undergoes rotational Brownian movement, until the active area is superposed on the active site, whereupon specific combination occurs. This second method, which we call *complex formation*, will be considered in more detail later.

Before continuing with numerical analysis, it must be pointed out that neither of the above include the idea of an activation energy, an energy barrier, which selects only those favored collisions in which random thermal energy exceeds a certain value. The presence of an activation energy, which is often observed, slows down the reaction, often drastically. It is our aim to examine what rates are possible, with the assumption of no activation energy. If, under the assumption, the processes supposed to occur are too slow, then we can be reasonably sure that the actual postulated mechanism is too slow and some change in thinking is called for. Of course, if the rates calculated for no activation energy are adequate, it is still no guarantee that the proposed biochemical process can perform the needs of the cell, until actual rates are available from experiment.

The validity of equation (2) can be tested by comparison with some measurements by Chance (1949) on the rate of formation of the enzyme-substrate complex between catalase and hydrogen peroxide, methyl peroxide and ethyl peroxide. The constants, (k_1), for this process are 3×10^7, $8{\cdot}5 \times 10^5$ and 2×10^4 liters per molecule second. Re-expressing as molecules per enzyme molecule and taking into account that in equation (2), C_0 is in molecules per cc if D and R are in cm^2/sec and cm, as is usual, we find

$$k_1 = 2\pi DRg \times 6{\cdot}03 \times 10^{20}$$

Since only approximate figures are of interest, we can take $D = kT/6\pi\eta R$ where k is Boltzmann's constant, T is absolute temperature and η is viscosity.

Then $k_1 = \dfrac{2kTg \times 10^{20}}{\eta}$ approximately, giving $k_1 = 1{\cdot}28 \times 10^8 g$ at 37°C in water. Thus for hydrogen peroxide, a small molecule, a value of g of 1/100 gives a rate which is too low. For the two others k_1 is too large, even with the small value of g. This comparison with experiment indicates that the rates calculated are not absurd and that the claim that this method yields plausibly rapid rates is, at least, fair. A very early analysis by Moelwyn-Hughes (1937) in terms of collision theory gave numbers which indicate about the same order of factor between collision rate and enzymatic rate.

Recent work by Strother & Ackerman (1961) tends to strengthen the reasoning. They studied the action of catalase on H_2O_2 in several media of

variable viscosity. For values of the viscosity ten times that of water the reaction rate followed the variation of the diffusion constant. They found agreement with the Smoluchowski equation, our equation (1), if R were 0·6 Å, which corresponds, in our treatment, to a value of g of 1/4.

In the interest of seeing whether wide discrepancies exist, we propose to continue this analysis with a value of 1/100 for g.

Using the theoretical value $D = kT/6\pi\eta R$ equation (2) becomes

$$\phi = \frac{kTgC_0}{3\eta} \tag{3}$$

or, at 37°C

$$\phi = \frac{1{\cdot}41 \times 10^{-14} gC_0}{\eta} \tag{4}$$

Collision Kinetics: Assumption of Immediate Correct Contact

Since the cell growth tells us the rates involved, equation (4) can be used to determine C_0. A reasonable value of a concentration, or at least one lower than is found, can be taken as evidence that the process is possible by Brownian motion and selection.

PROTEIN SYNTHESIS

We can assume the process for analysis is:

[1] (Enzyme ATP) + AA → Activated AA or "AAA"
[2] AAA + SRNA → SRNA — AA
[3] SRNA — AA → Template on ribosome → protein (4).

Here we use AA for amino acid and SRNA for soluble RNA. Taking 5 amino acids of one kind processed per second and analysing processes [1] and [3] we obtain, for $\eta = 1{\cdot}0$ poise.

$$5 = 1{\cdot}41 \times 10^{-16} C_0$$

Then $C_0 = 3{\cdot}5 \times 10^{16}$ molecules per cc
$= 3{\cdot}5 \times 10^{4}$ per cell, assuming 10^{-12} cc is the available volume of the cell. For 20 amino acids, this requires a total pool of $7{\cdot}0 \times 10^{5}$ per cell. This is about as many amino acids as are incorporated per second, and does not seem unreasonable.

The same figure holds for SRN–AAA. A pool of $7{\cdot}0 \times 10^{5}$ SRN–AAA, of molecular weight 25,000 each, is a total mass of $2{\cdot}9 \times 10^{-14}$ gram, compared to $4{\cdot}4 \times 10^{-14}$ gram total RNA, or roughly 70% of the total SRNA. This is a very high figure. If we include process [2] above, involving the combination of SRNA and activated amino acids, then even

more SRNA is needed. Thus *if* the type of mechanism discussed is operating, the cell is placing excessive demands on SRNA. This suggests that either some other form of collision kinetics be used, or that some attention be paid to the way in which SRNA actually operates.

DNA SYNTHESIS

For one point of DNA synthesis ϕ is 3×10^3 per second for one kind of base. Then $C_0 = 2{\cdot}1 \times 10^{19}$ or $2{\cdot}1 \times 10^7$ per cell and for four bases, roughly 10^8 per cell. A mass of $0{\cdot}75 \times 10^{-14}$ gram of DNA contains $1{\cdot}5 \times 10^7$ nucleotides. The figure required for C_0 is far too high. Again this suggests that if this mechanism is valid, there must be not less than 100 separate points of DNA synthesis.

FORMATION OF RIBOSOMES

It is unwise to devote much time to theory on the formation of ribosomes until the factors that concern their interrelationships have been established. In a bacterial cell there are several kinds of ribosome and they are in some kind of dynamic equilibrium. We can examine one very simple suggestion, namely that the ribosome RNA is all "made" separately from the protein and that it is formed into ribosomes by specific combination with protein, made elsewhere, but specifically related to the RNA.

Since there are 10^4 ribosomes per cell, we can use 10 per second as the rate of formation of ribosomes. If we further suppose that 10 protein molecules must find and combine with each unit of the RNA, we have a definite process to which equation (4) can be applied. We can suppose that 10 RNA molecules are made per second and inquire as to the concentration of protein necessary to prevent the combination from slowing down the "processing" of ribosomes. Since each RNA molecule must combine with 10 protein molecules, the rate of individual combination of protein molecules is 10 per second.

Then, by equation (4), $C_0 = 7 \times 10^{14}/g$ or $7 \times 10^2/g$ protein molecules per cell must be available for combination with RNA. The total number of protein molecules per cell is, in round numbers, 10^6, so that a fraction of 1/1,000 which are in process is reasonable. Such a process requires g to be nearly unity, or that the combination be almost nonspecific. If the normal value of 1/100 is taken for g, then 7% of the total cell protein must exist as "pool" for ribosome synthesis. The impression we form from these figures is that either protein–RNA combination in ribosomes is not very specific, or that other factors exist; such as a means for speeding up specific combination, or a method of formation of ribosomes in which the diffusion distances are far less.

Collision Kinetics with Complex Formation

The previous analysis leaves the definite impression that more than one of the synthetic processes are marginal if all that takes place is collision due to Brownian movement with specific capture at a favored collision which presents the right aspect. The second approach, involving the formation of a complex, previously outlined, introduces an additional degree of freedom to the process and removes some of the restriction on the nature of a successful collision. The London–Van der Waals force, due to transient dipoles producing induced dipoles, causes an attraction which continues in spite of rotation of the colliding molecule. Therefore a collision in which a molecule enters with the range of attraction of a larger molecule surface, may cause the colliding molecule to remain near the specific surface long enough for rotational Brownian movement to cause the active part of the molecule to approach the synthetic site. The kinetics of this process are different and, on the whole, more favorable.

ROTATIONAL BROWNIAN MOVEMENT

By analogy with the equation $D = kT/6\pi\eta R$, a rotational diffusion constant Θ can be used to describe the amount of rotation developed by a colloidal particle in a given time. The mean square angle traversed, θ^2, is given by $\theta^2 = 2\Theta t$, where $\Theta = kT/8\pi\eta R^3$ (see, for example, Alexander & Johnson, 1949).

If we set t_1 as the time for the square root of the mean square angle to be 2π then t_1 is an estimate of how long it will take for a favorable aspect to be presented. For an amino acid of radius 3·5 Å t_1 is $5{\cdot}0 \times 10^{-7}$ second; for SRNA, of radius 16·5 Å, the time is 100 times longer, or $5{\cdot}0 \times 10^{-5}$ second. Both times are very short and it is clear that if a complex can form, which can last for as short a time as a microsecond, then any specific combination will have an excellent chance of developing.

The difficulty in this analysis is obviously in the nature of the complex. We can obtain some insight into what may be occurring by recalling that Vold (1954) found that for distances of surface separation less than the cube root of the volume of colliding particles the attractive force is of the order of kT, and increases for lesser distances. In our formulation the cube root of the volume is best represented by R. Now suppose the value of the angle traversed while linear Brownian movement carries the molecule through a distance R is examined. This figure corresponds to the angle described within the attractive field assuming no binding at all. Any binding, however transient, will increase the time for angular Brownian movement and so will increase the figure obtained.

Using the two relations $X^2 = 2Dt$

and $\theta^2 = 2\Theta t$

where x^2 is the mean square distance gone by linear Brownian movement; and the values for D and Θ, we find, if we set $x = R$

$$R^2 = \frac{2kT}{6\pi\eta R}\, t$$

$$\theta^2 = \frac{2kT}{8\pi\eta R^3}\, t$$

$$\text{or } \theta^2 = \tfrac{3}{4}.$$

This rather interesting result shows that approximately one radian of rotation occurs within the influence of London attraction. This is $1/4\pi$ of the total needed for the whole surface, since each angle, "azimuth" and "elevation" is equally variable. Thus there is no reason to assume a g value of less than $1/12$, if the rotation of the molecule is an important factor. If a complex forms, of even a very weak and transient nature then it would be expected that g values of unity would be common.

There is no need to repeat the calculations of the previous section. The conclusion reached is that *if* rotation and transient binding are possible, there are no synthetic processes requiring undue concentrations. It would even be possible for the DNA to be synthesized at one single point, though only marginally.

Precision of DNA Replication

The most precise process so far discovered in the operation of the cell is the duplication of DNA. Recent work on the nature of mutation, notably that of Gierer & Mundry (1958) indicates that the incorrect location of one base can be expressed as a mutation. The work of Novick & Szilard (1950), for example, has shown that the frequency of mutation to resistance to T5 phage is 10^{-8} per 50 minute generation time. If it is assumed that this property is determined by a length of 1,000 nucleotides, in accordance with estimates for the length of a cistron, then an error anywhere among these nucleotides can be observed as a mutation, meaning that one false assembly in 10^{11} assemblies per generation time is all that takes place. This figure is most important in attempting to analyse the rates of biological synthesis. A systematic search for the most accurate assembly process, meaning the least mutation rate in the longest cistron, would be worth while. Certainly many events which are, at present, treated as mutations, may be due to some other type of process than that described. Since this a theoretical analysis, we state our premises. It is supposed that the false assembly of any one nucleotide in a cistron of 1,000 nucleotides

appears as a mutation and that a possible rate of such mutation is that measured by Novick & Szilard at 10^{-8}.

It is proposed to examine whether such a precision can be attained if the cell operates so that *precision in time is important*. If it is, then an idea of the rate of DNA synthesis at one synthetic point is important. If only *one* such point exists, which was indicated to be marginally possible, then the rate of synthesis is $1{\cdot}2 \times 10^4$ per second. With the assumption that precision in time is important, there is required to be an assembly, by random bombardment and specific selection, of 10^4 per second, in round numbers, such that each is made with a chance of error of 1 in 10^{11}. Since the collision process is random, the specific selection must change with precision to assemble the correct bases. The assumption made is that the precision is achieved by rapid change of the process selecting the base and that the chance of error is expressed in terms of the finite rate of change. Since this is a rather special requirement an analogy may be helpful. If a set of code letters were presented on a screen and the process of transmitting the code consisted of a series of rapid, but randomly timed photographs of the screen, then one factor in the correct transmission of the code would be the rate at which a letter was replaced, because any picture taken during the change would amount to an error. It is not expected that the cell actually performs DNA synthesis in this way; it is the purpose of this discussion to examine *whether it could*, and if not, by what margin.

Under this assumption, then, the error in time may not exceed $10^{-11}/10^4$ or 10^{-15} second. This is short enough to suggest examination in terms of the uncertainty principle. If the required uncertainty in energy is ΔE to accompany an uncertainty in time of Δt, then

$$\Delta E.\Delta t = h/2\pi$$

where h is Planck's constant. Setting $\Delta t = 10^{-15}$ second, we find $\Delta E = 10^{-12}$ erg, or 0·6 electron volt. This is a very large line width for a process of stable binding and suggests that precision in time is *not* important. Even if 100 sites of synthesis were operating which increases the available error in time by a factor of 100, the value of ΔE is still surprisingly large.

Even with a time as long as 10^{-13} second, involving 100 sites of synthesis, there could not possibly be physical motion of the enzyme or the primer from one base to the next. This distance is 3·4 Å and if we take an enzyme of molecular weight 100,000, the velocity of motion between bases has to be $3{\cdot}4 \times 10^{-8}/10^{-13}$ or $3{\cdot}4 \times 10^5$ cm/sec. This yields a kinetic energy of 6,000 electron volts—far too high.

These considerations seem to eliminate precision in timing as the means by which precision in base order is obtained. Many quite reasonable alternatives present themselves. One is that the "enzyme" is actually a

complex of a protein molecule, of rather low specificity, and the primer. The active synthetic site is at the point where the new chain ends and the next base lies unpaired. The new base is then automatically selected, regardless of time, as soon as the pyrophosphate has diffused away. The selection is also fast enough to permit the observed rates.

Another possibility, which retains motion of the enzyme, is that it loses all synthetic function while moving from one base to the next, and only becomes functional when a new, firm combination with the new base on the primer takes place. Then the time to transfer is unimportant. Both mechanisms suggest definite properties for the enzyme which are worth investigating.

Transfer from Rotation to Vibration

The analysis of synthetic rates just given shows that if only part of a molecule is active, and that this part must approach closely to a synthetic region, then a direct collision process tends to give rates which are rather low. On the other hand, a process in which a very temporary complex forms, permitting rotation to occur, and so allowing selection of the active region, is about 100 times faster.

For this second process to operate there must be reasonably free rotation of small and medium molecules, which means they possess rotational energy and angular momentum. On the other hand, the very nature of synthesis means that rotation of a small molecule has ceased. Thus there has to be a transfer of rotational energy to vibrational energy, and a change of angular momentum. The effect of a considerable difference between the free and the bound state would be to diminish the number of successful collisions. To take an extreme case, if the lowest energy of binding were above the thermal energy of rotation, then only those molecules favored by abnormal energies could combine—essentially the requirement of an activation energy. The same restriction could apply to a change of angular momentum. This suggests that one factor in a synthetic chain of events may be the transfer of angular momentum. Thus one function for SRNA could be to permit a capture of an amino acid into a bonded state with some free rotation and thus make the transfer to the nonrotating condition in the protein chain a two step, rather than a one step, process. The successful incorporation of analogs may also depend, in part, on their ability to allow change of angular momentum.

Conclusion

The effectiveness of what are sometimes called "diffusion limited" reactions in several synthetic processes has been examined. If activation energies are low, or not present, then the observed rates can be matched by

expected rates, though without a tremendous excess. Any considerable activation energy would have to be confined to relatively nonspecific reactions involving small metabolites. It would be expected that an enzyme process requiring relatively large activation energy could only be possible with many enzyme molecules and a high substrate concentration. Such enzymes, in high concentration, would be easy to isolate. Their study may give a false picture of the actual workings of the cell.

Measurements on cellular viscosity are very much needed. Also studies of enzyme kinetics in high viscosity media should be made, along the lines of the work of Strother & Ackerman. Attention to the physical meaning of "specificity" is important. Factors not even discussed here may be operating, and may need the use of quantum biochemistry to elucidate them.

The precision replication of DNA by this very crude assembly line technique, bears close examination as to rates and permissible errors. Most genetic studies are essentially static. It may be that genetic dynamics is more revealing of the processes of life just as dynamics is more revealing of the physical world than statics.

Several discussions with Dr. G. K. Strother were very valuable.

REFERENCES

Alexander, A. E. & Johnson, P. (1949). "Colloid Science", 384. Oxford University Press, London.

Bessman, M. J., Lehman, J. R., Simms, E. S. & Kornberg, A. (1958). *J. biol. Chem.* **233,** 171.

Chance, B. (1949). *J. biol. Chem.* **179,** 1341.

Chandrasekhar, S. (1943). *Rev. mod. Phys.* **15,** 1.

Collins, F. C. & Kimball, G. E. (1949). *J. Colloid Sci.* 425.

Crick, F. H. C. (1958). In "Symposia of the Society for Experimental Biology", No. XII, 138. Cambridge University Press, London.

Dintzis, H. M., Borsook, H. & Vinograd, J. (1959). In "Microsomal Particles and Protein Synthesis". Pergamon Press, London and New York.

Gierer, A. & Mundry, K. W. (1958). *Nature, Lond.* **182,** 1457.

Hanawalt, P. (1958). Ph.D. dissertation, Yale University.

Hoagland, M. B., Keller, E. B. & Zamecnik, P. C. (1956). *J. biol. Chem.* **218,** 345.

McQuillen, K., Roberts, R. B. & Britten, R. J. (1960). *Proc. nat. Acad. Sci. Wash.* **45,** 1437.

Moelwyn-Hughes, E. A. (1937). In "Ergebnisse der Enzymforschung", VI, 23. (Nord, F. F. & Wiedenhagen, R., Eds.) Akadem Verlagsgesellschaft, Leipsig.

Novick, A. & Szilard, L. (1950). *Proc. nat. Acad. Sci. Wash.* **36,** 708.

Overbeek, J. Th. G. (1952). In "Colloid Science". (Kruyt, H. R., Ed.) Elsevier, New York.

Quastler, H. (1953). In "Information Theory in Biology", 41. University of Illinois Press.

Smoluchowski, M. von (1916). *Phys. Z.* **17,** 557, 585.

Strother, G. K. & Ackerman, (1961). E. *Biochim. Biophys. Acta.* In press.

Vold, Marjorie J. (1954). *J. Colloid Sci.* **9,** 451.

Zamecnik, P. C., Stephenson, M. L. & Hecht, L. I. (1958). *Proc. nat. Acad. Sci. Wash.* **44,** 73.

J. Theoret. Biol. (1961) **1**, 342–358

Properties of Large Molecules that go beyond the Properties of their Chemical Sub-Groups

John R. Platt

Department of Physics, University of Chicago, Illinois, U.S.A.

(*Received* 30 *January* 1961)

Kasha has listed some of the physical and chemical properties that polyatomic molecules of 5 or more atoms begin to have that might not have been anticipated from a knowledge of diatomics and triatomics alone. Here a list is given of properties that "fully-complex" molecules of 500 or more atoms may begin to have that go beyond the collection of properties of their chemical sub-groups or the properties of smaller systems.

The list includes mechano-chemical properties such as contractility and contractive forces in long flexible chains, and reproducibility in successive amplification processes of catalysis or response. It includes the effects of molecular continuity in facilitating sequential chemical reactions; the specificity of information and of reactions that can be obtained with complex sequences of side groups; long-range order; polyelectrolyte behavior; and other cooperative properties. Several of these properties are related to the action of muscle. Because the various properties appear necessarily and naturally with increasing molecular size, they make it possible to explain a number of strange processes in biology more easily in ordinary physical and chemical terms without invoking additional peculiarly biological assumptions.

1. Changing Approximations and Changing Properties in Different Ranges of Molecular Size

It is not always realized how much the most interesting and important properties of molecules depend upon their size. Kasha has emphasized this by giving an instructive list of some properties that polyatomic molecules have that might not have been anticipated from a knowledge of diatomics and triatomics alone (Kasha, in preparation; see also Table 1). The present paper is concerned with some of the further properties of macromolecules that might not have been anticipated from a knowledge of the properties of their polyatomic chemical sub-groups or the properties of smaller molecules.

A. Changes of Physics with Changes of Scale

We usually say, of course, that the fundamental principles of physics and chemistry do not change with a change in molecular dimensions. But

this is only a theoretical dictum, and when it comes to actual quantitative predictions or calculations of relationships, we generally use a different set of rules or equations for problems in different ranges of size. The basic laws may not alter, but they are combined in different proportions, so to speak; so that the correction terms change, the best approximations, change, and consequently the dominant physical forces and the distinctive properties of matter change, with every alteration in scale.

Small water waves, for example, are governed principally by surface tension, while large ones are governed by gravity, although both forces are always present. Atoms and small molecules must be described by quantum mechanics, but this is in many ways approximately like classical mechanics, as Bohr's "correspondence principle" asserts, and it goes over more and more exactly into classical mechanics in describing the macroscopic properties of large molecules.

In biology, where the pressure of survival can lead to great amplification of any small "correction terms" useful in the maximization of performance, the complete change of structure and mechanism with every change of scale is particularly striking. The tiniest creatures float in air; larger ones buzz; still larger ones flutter; at the next size, they fly; then they soar; and finally, for the largest creatures, aerial support becomes impossible (Thompson, 1959). Each order-of-magnitude change in the weight makes a distinct change in the way the mechanism operates. (See also Bertalanffy, 1952.)

B. THE RELATIVITY OF APPROXIMATIONS

Even in describing atoms and small molecules, our "fundamental principles", whatever their universal validity in theory, are in practice only approximations that are more or less appropriate to a particular range of problems. The fundamental Dirac equation for the electron is an approximation that omits the vacuum polarization; the Schrödinger equation, which is usually taken as the basis of chemical quantum mechanics, omits relativity; the Hartree-Fock equation omits electron readjustment and correlation energies; the "simple LCAO" equation that we use for conjugated molecules omits all but the pi-electrons; and so on. The larger the molecule, the more steeply complex the equations, and the cruder the approximation must be if any practical computations are to be made at all. Yet each approximation in turn becomes too complex to be workable, or its accumulated errors begin to make it unreliable, for some still larger system.

Thus the Schrödinger equation cannot be solved rigorously without further approximations for any chemical problem larger than the 3-particle problem of the H_2^+ one-electron bond (Kolos, Roothaan & Sack,

1960). And even if it could, its errors from neglect of relativity would reach hundreds of electron-volts for, say, the inner-shell electrons of the Fe atom, a quantity large enough that even small fluctuations in it might conceivably have appreciable effects on chemical binding. The Hartree-Fock equations probably cannot be solved for asymmetrical systems of more than about a dozen atoms. And the usual equations that neglect retarded-potential in electron-interactions become badly in error for systems absorbing at wavelengths beyond about 1 micron (Platt, 1960); and so on. Faster computing machines and new theoretical insights may lead to larger practical ranges of validity, but they will not eliminate the phenomenon; and the relativity of approximations will probably always be a central feature of physical science.

This is why the different scientific disciplines will never completely merge. It is extremely fruitful for us to go as far as we can in explaining biology in terms of biochemistry, biochemistry in terms of chemistry, and chemistry in terms of physics and quantum mechanics. But men of the greatest experience and insight cannot be perfectly certain of their approximations when they extend their results from one field into another. Major differences in their equations may become smoothed into constant and therefore negligible background effects, while minor correction terms may be elevated into differential effects of major importance for prediction. The success of their intuitions and approximations obviously cannot be judged against an uncomputable absolute standard, so it can only be checked and corrected by empirical reference.

This means that every science that is a science will always have to develop its own peculiar and powerful methods of inference and methods of organizing and structuring its field. It cannot abdicate the search for its own rules and its own symbolic representations of its own empirical conclusions, just on the prospect that, with infinitely elaborate computations and great mathematical insight to pluck out the results, it might someday be quantitatively derivable from some more fundamental discipline. Its own rules—its own "new" properties of matter—are in fact the best guidance to the theorist as to what the appropriate approximations are. We see that the independence of the disciplines is just as necessary to a healthy science as their interrelationship.

It is surprising how often these rules of fruitful interaction between the different areas of study are ignored. Molecular orbital theory, with its many-dimensional equations, has often tended to ignore the three-dimensional invariances represented by the chemist's valence bond, which has been the central fact of chemistry for a hundred years (Platt, 1961). Hundreds of brilliant man-years may have been side-tracked in molecular calculations with inappropriate approximations. Likewise for

fifteen years, some of the best infrared theorists disbelieved in the possibility of separable group frequencies in infrared molecular vibrations, even when these frequencies were being used for analysis daily in the laboratory.

The reason for describing in such detail these difficulties of theoretical approximation is that we need to keep them in mind when trying to go on to anticipate how molecular properties may change further in going to a new range of sizes. It is well to be skeptical of dubious results, but it is also well to be prepared for real surprises and for the appearance of new regularities in the properties. We should therefore keep ourselves receptive to any reliable biochemical and biological evidence that will help guide and correct our guesses.

C. CHANGE OF MOLECULAR COMPLEXITY WITH CHANGE OF SCALE

Broadly, there are two groups of changes in the physical and chemical properties of large molecules that occur naturally and necessarily with every change of scale. The first group consists of the expected changes, for example in vapor pressure or viscosity, that result simply from the difference in size or weight. Our emphasis here, however, will be on the second group, the unexpected changes represented by major differences in important principles of behavior. How can such differences develop? We discover on thinking about them for a while, that they develop principally from the fundamental increase of possible complexity as molecules get larger. The atom is an absolute unit; and large molecules do not differ from small ones merely by a similarity transformation, but by containing more atoms, so that more complex behavior is possible.

This increased complexity often surprises the physicist when he first comes upon it, because the many-atom systems that he is familiar with usually do not show such a size-dependent complexity. His perplexity has been discussed recently in an elegant lecture by Burton (delivered at the Conference for Teachers of Biophysics held in Miami last year) who shows that it is due to the fact that there are actually three quite different kinds of many-atom systems. These are: (1) systems with a simple, or perfect, repeating pattern, such as perfect crystals; (2) systems with an imperfect pattern such as liquids, or crystals containing dislocations or impurities, where most of the variations or imperfections are random or "meaningless" and can be treated successfully by statistical averages. The detailed patterns are never copied by being duplicated or amplified. And (3) systems with an organized or non-random complexity in which each atom or group has a specifiable role, as in the more complex molecules of organic chemistry or biochemistry. The "role" appears when the detailed patterns are copied or their detailed arrangement affects an amplification process.

Burton points out that systems of types (1) and (2) are the kind that the physicist has usually worked on; while it is those of type (3), the "fully complex" systems, that are the ones important to the biologist. What is more significant for our purpose here, the systems of types (1) and (2) are those whose properties are essentially invariant to a change of scale, once enough atoms are present to reduce the "end effects", because the details are only significant statistically, as averages. But systems of type (3) can continue to change their properties radically with every change of scale, because their complexity of organization changes radically. In fact one might say that this is the major difference between biology and physics. The properties of many-atom systems that we want to examine here are not the somewhat repetitive properties of the perfect or random systems, but the size-dependent properties of these "fully complex" molecules.

Kasha has begun this discussion of how molecular properties change with molecular size, by giving us the list in his Table 1 of the important

TABLE 1

Properties of molecules in the 5- to 50-atom range that have no counterpart in diatomics and many triatomics

(After Kasha, private communication)

"Obvious" geometrical properties

1. Angular dependence of bonds, bond forces and reaction paths (beginning with 3 atoms)
2. Anisotropic absorption (distinguishable x, y, and z axes—beginning with 3 atoms)
3. Internal rotation (beginning with 4 atoms in a chain of 3 bonds)
4. Stereochemical properties (beginning with 4 non-planar, non-identical atoms)
5. Optical rotation and magnetic anisotropy (from (4))

Less obvious properties

6. Jahn-Teller effect of electronic-vibrational interaction in degenerate states of symmetrical non-linear systems (beginning with 3 atoms in triangular arrangement)
7. "Internal conversion" of excitation energy (beginning with polyatomic molecules in solution)
8. Long-lived excited triplet states (from 7)
9. Formation of micelles and surface films by hydrophobic-hydrophilic molecules in solution (beginning with about C_{10} fatty acid chains)

Properties of electron-deficient or conjugated heavy-atom networks
(hydroborons, polyenes, aromatics, porphyrins, etc.)

10. Non-additivity of atomic properties (beginning with butadienes)
11. Non-equivalence of similar atoms (beginning with C_5 hydrocarbon chains or rings)
12. Chemical "directing effects" and interactions of distant substituents (first prominent in benzenes)
13. Electron "mediation" through conjugated systems, facilitating catalysis of electron-transfer (Taube) reactions
14. Strong electronic absorption in the visible and solar region (first prominent for common conjugated systems in C_{10}–C_{14} chains or aromatics or heteroaromatics)
15. Charge-transfer forces and formation of charge-transfer complexes or pi-complexes

physical and chemical properties that 5- to 50-atom molecules begin to have, which have essentially no counterpart in diatomics or sometimes in triatomics or even larger systems. There is only one property which may surprise many chemists and biologists by its omission, namely, energy-transfer between polyatomic molecules after excitation; but this property, so important now in the molecular case, was known already for atoms from the classical resonance experiments of Franck and Cario, and therefore does not belong in such a Table.

Here we go on to ask what are the further properties that become possible in molecules an order of magnitude larger, in the 50- to 500-atom range of sizes and up. The properties we shall discuss are listed in our Table 2. They represent an extrapolation from known properties of

TABLE 2

Properties of molecules in the 50- to 500-atom range and up that go beyond the properties of their chemical sub-groups

Long-chain properties

Properties arising from connectedness

1. Possible contractility (from multiple internal rotations)
2. Variable viscosity in solution (from 1)
3. Mechano-chemical contractive forces, solvent-dependent or reaction-dependent, between points on the chain (from 1)
4. Long-range specific and directed interactions between groups attached to the chain (from 3)
5. Stability and reproducibility in successive amplification processes such as catalysis and response

Properties arising from continuity

6. Ease of repetitive construction and repetitive operation in reactions
7. Reduction of entropy factors in "hand-to-hand" reactions along the chain

Properties arising from sequential complexity

8. Specificity of sequence, multiplying with chain length
9. Self-replication (from 8 for chemically reacting chain with a sequence longer than some unknown minimum size)

Other cooperative properties

10. Long-range order (secondary; tertiary; and long-range "crystallization")
11. Polyelectrolyte behavior.
12. Possible "wrap-around" cooperative effects on binding energies and activation barriers
13. Possible selective transfer of electrons through large ions and radicals

smaller systems, but they have been considered with the macromolecular evidence in mind, and it is certainly responsible for suggesting a few of the more surprising results. Any such list is tentative, of course, because of the difficulties we have already noted in trying to anticipate correctly all the new phenomena in a new range of sizes. No doubt further consideration will show other peculiar properties that will add to or change or contradict these present inferences.

The kinds of molecules emphasized in Table 2 are seen to be somewhat different from those in Kasha's Table 1. For one thing, we need not go on enlarging the conjugated systems. Apparently the largest ones that occur in biology are carotenes and porphyrins, in the range of 20 to 30 conjugated atoms. These chromophores occur close-packed in some biological structures such as chloroplasts, but they are not conjugated to each other; their spectra and other properties are essentially additive and there is no evidence that chemical conjugation effects or directing effects are propagated further than the length of one individual chromophore. Larger conjugated systems are often more unstable, and probably they absorb unnecessarily far in the infrared. In addition, perhaps they are too much like microcrystals of graphite, with rather repetitious reactions and no internal motions, so that they may not be "fully complex" enough for further biological use.

On the other hand, we do need to emphasize the special properties of large chain systems that have strong covalent bonds capable of internal rotation, because of the present evidence that polypeptide and polynucleotide chains of this type dominate the macromolecular scene. Indeed a little reflection shows us that one-dimensional linear chain molecules which may have a large variety of configurations in space can be more fully complex in their properties and behavior than two-dimensional molecules. Threads or ropes, as we all know, can be twisted or tied into thousands of different knots or knitted into sweaters (and pulled out again!); while sheets can only be flat or folded or crumpled, if they are "developable" surfaces, and have even less freedom if they are not. A three-dimensional structure has almost no configurational alternatives and so has no complexity of behavior beyond its original structural complexity, except to the extent that it may be made of a porous and flexible network of threads whose sections have one-dimensional freedom. We begin to see why linear chains play such a central role in biology. It is evidently convenient to divide the new properties of large and fully-complex molecules into two classes, as shown in Table 2; first, the properties associated with long and flexible chains; and second, the other cooperative properties less dependent on topology. Let us now look at these properties one by one.

2. Long-Chain Properties

A. PROPERTIES ARISING FROM LINEAR CONNECTEDNESS WITH STRONG LINKS

1. *Contractility*

A three-dimensional network has no degrees of freedom for rearrangement of its structural relationships without breaking bonds. A two-dimensional sheet network offers one degree of folding at each point, but

it has many internal constraints, so that a family of points parallel to the fold are constrained to fold together. A one-dimensional thread offers the possibility of two degrees of rotational freedom at each point. (In a simple chemical chain, there may be only one degree of rotational freedom about each single bond, since the bond angles are fixed; but the groups at the ends of two non-linear adjacent bonds will have two degrees of rotational freedom with respect to each other.)

Internal rotation about a few bonds will produce little change in the dimensions or properties of small and compact molecules. The *cis-trans* changes in polyenes are among the largest effects of this sort in the 5- to 50-atom range. But in a 500-atom or larger system, a few internal rotations in the right places can produce order-of-magnitude changes in the dimensions. A pH or solvent change or a chemical reaction which acts either on the rotation angles of many of the bonds or on the binding of a long molecule to itself rather than to the solvent, may produce a change from a maximally extended form to a maximally contracted form.

A 50-atom polypeptide containing, say, 4 amino-acid side groups on a 10-atom chain, has almost the same dimensions in every possible configuration. But a 500-atom polypeptide chain can change its greatest dimension by about a factor of 3 from its most extended to its most compact form. The effect is greater for chains than for sheets, and is greater, the greater the size of the system. A double-helix chain of bacterial DNA, say 1 cm long and 30 Å in diameter, would have its greatest dimension reduced by a factor of 10^4 if it were packed into the smallest possible volume.

2. *Variable viscosity of solutions of chains*

This is a consequence of 1.

3. *Contractile forces between points on the chain*

It is not so commonly emphasized in molecular biology that the possibility of contractile forces is necessarily present in any system that can have a variable configuration or that shows a variable viscosity. We are familiar with many side aspects of chemical reactions, those, for example, that we study under the names of "surface chemistry", "electrochemistry", and so on. But a somewhat neglected side-effect, and one probably of great importance in long-chain chemistry, is what might be called "mechano-chemistry". This would be the study of the mechanical swellings and contractions and rope-like motions, with the corresponding forces, that can result naturally from chemical reactions in a connected chain system.

Numerous cellular processes, such as protoplasmic streaming, spindle

formation during mitosis, and the interleaved contraction of actin and myosin units in current theories of muscle action, appear to involve the mechano-chemistry and forces of long-chain molecules expanding and contracting. It is important for the biochemist to begin to realize that there may be many such phenomena that are not necessarily peculiar to *in vivo* systems, but that can probably be found and studied *in vitro* whenever we have a long connected chain of reactant groups. The actin-myosin contraction with ATP is of course an important demonstration of this kind even though little is known as yet about the molecular configurations during the process.

4. *Long-range interactions*

Such contractile forces can be long-range, directed and specific. Their range can be any range up to the maximum length of the chain molecule. Any two groups on the chain can be brought toward each other, or even together if the chain wraps around itself. And the specificity can be as great as the specificity of the sections of the chain; the specificity perhaps being that of the reactant groups on the chain, or perhaps that of external reactant groups attached or complexed to them. This kind of long-range force in biology could be of much longer range and much more sharply directed than any of the disputed "long-range interactions through space" that have been postulated by some writers on biology.

Some detailed "wrap-around" mechanisms for helical chain motions during or between replications have been explored earlier (Platt, 1955).

5. *Stability and reproducibility in amplification processes*

"Selective response" and "enzyme action" or "catalysis" are names for biological amplification processes which convert a small selective input stimulus into an amplified output response. In order for such responses or reactions to be reproducible time after time, it is necessary to have a stable amplifier, that is, an amplification system that comes back to its original configuration no matter what intermediate forms it may assume (Platt, 1956).

Obviously a collection of small molecules with statistically varying arrangements will hardly satisfy this requirement. A selective amplifier needs a fairly complex structure (Platt, 1956). A large three-dimensionally connected system can have structure, but will be relatively inflexible, so that it is a catalyst either all the time or none. It is like a photocell, which responds (except for statistical losses) every time a photon hits it; so it offers little possibility of a time-dependent internal control, that can turn the response on or off at one time or another by means of internal changes. A large two-dimensional structure offers more flexibility, but a one-

dimensional structure offers the most flexibility of all in terms of alternative time-dependent configurations which do not alter its primary sequential structure. A chain may therefore be a particularly favorable way to construct a time-dependent but reliable amplifier.

This amplifier language is only another way of saying what Schrödinger (1956) said when he originally emphasized the importance of tightly bonded structures in explaining the stability and reproducibility of life processes. He emphasized that thermal fluctuations would disrupt the weakly bonded arrangements of small molecules, and that large molecules were needed, with strong covalent bonds "stabilized by quantum-mechanical principles", in order for the activation barriers to be high enough to prevent rapid loss of the hereditary information by thermal disruption.

B. PROPERTIES ARISING FROM CONTINUITY

6. *Ease of repetitive chemical operations in construction or functioning of long continuous chains*

A long chain is a complex product, but it is one that may be, in some ways, comparatively easy to produce chemically. The temporal continuity of any stable chemical flow process, such as the complex flow through living systems, implies repetitive processes. Such processes may produce small and separate repetitive products or they may produce continuous chains with a repetitive backbone, as we see in the laboratory in the serial growth of synthetic polymers in solution, or in the growth of a synthetic rubber chain from a reaction site on a catalyst. Such chains may grow indefinitely, their termination being accidental and unrelated to the growth mechanism. Large structural chains or rods may be grown from a small and simple synthetic site. And this simplicity of repetitive synthesis may extend in some degree to repetitive backbones even when they bear an organized and non-repetitive sequence of side-groups.

Possibly a continuous chain may not only be made by serial chemical operation but may also be able to carry out serial chemical operations comparatively easily. If it contains a series of reactants, their continuous connection might facilitate the carrying out of a series of consecutive reactions on a changing reaction-intermediate molecule as it passes in small steps from one reaction site to the next. Koshland (1958) and others have suggested such consecutive reactions along a chain in forming a specific protein from a template.

7. *Continuity and the reduction of entropy factors*

Such chain processes, in which reactions would proceed stepwise down a chain or in which two reactants would be brought together by long-range directed contractions (as discussed in 4), would have a second im-

portant feature. This is the reduction of entropy factors by the elimination of diffusion-controlled steps, with a corresponding increase in the speed and directness and yield of each reaction. The reduction in diffusional delays and in the losses and misdirection of intermediates would reduce the amounts of reactants needed, a matter of great importance when there is a long series of reactions involved.

It is believed that the time required for diffusional processes is not a limiting factor in biochemical syntheses. But even small diffusional losses in yield at successive steps would be cumulative and serious in many syntheses, so that the yields may be a more sensitive test for "continuous attachment" to the intermediates between steps than the times would be. Certainly many biochemical syntheses proceed through a series of steps with such a high yield that it is hard to believe that there is usually any diffusional loss of reaction intermediates at any step. Chain processes, in which the intermediate is passed "hand to hand" from one reaction site to the next down a chain, would be a simple explanation of the "continuous attachment" that these yields therefore seem to imply.

How long might such "hand to hand" chains of reactants be? There is some evidence that in simple organisms the DNA in a cell may be a single long molecule, of the order of 1 cm long in some bacteria. This leads to the question whether many of the enzymes and other proteins in the cell, formed by instruction from the DNA, may not be similarly linked together in the cell into a chain of comparable length, "all one string". The breaking up of such a string into the small and sharply characterized fragments that we see and identify outside the cell might then be simply a rather reproducible artifact of our extraction and purification and characterization procedures.

There is empirical evidence for at least a certain amount of such linking, in the fact that successive genes on the genetic map (DNA), stretching through a linear distance of the order of microns, frequently control successive steps in a specific synthesis, as Suskind (in press) found for 6 of the 7 steps of the histidine synthesis in bacteria. This certainly suggests that the reaction intermediate in the synthesis may be proceeding in sequence "down a chain" of successive and complex enzymatic sites. At least such a supposition seems simpler than the supposition that the intermediate must break away after each step and then must try various reaction sites, locating the next proper one by diffusional trial and error. On the other hand, if chain connection is the general rule, an occasional folding back of the chains will then have to be postulated in addition, to account for the occasional cases where the reaction sequence is known to differ from the genetic sequence. (See also Demerec & Hartman, 1959; and Ames, Garry & Herzenberg, 1960.)

What would be the mechanical, or the mechano-chemical, sequence of molecular configurations in such a passage of an intermediate down a chain of reactants? One may perhaps imagine a "wave of contraction" passing down the chain, somewhat like the wave of motion in the progress of an earthworm. With "continuous attachment", the reaction intermediate might be "carried along" with the contraction, being transferred from one reaction site to the next as the contraction brings the sites close together and as each step of the reaction is completed (Koshland, 1958). The reaction energy might be supplied by numerous diffusible small molecules such as ATP in the surrounding medium. The eventual reaction product might be a small terminated product like histidine, or it might be a second chain as long as the reactant chain and with a structure in some way complementary.

Regardless of the accuracy of such configurational conjectures, it may be worth pointing out that the viscous losses of energy in longitudinal expansion-contraction waves in a chain might be extremely small compared to the losses in any transverse motion (other than Brownian motion). Thus there might be a strong tendency in chain systems for motions of any kind and for reaction paths of any kind to be physically parallel to the chain directions, like the tendency of an ice skate to move parallel to its own length, in the direction in which the resistance is less.

C. PROPERTIES ARISING FROM SEQUENTIAL COMPLEXITY

8. *Specificity of sequence*

It is now common to think of a sequence of chemical reactants on a chain as a "code" carrying the "information" that determines chemical specificity. All that we need to add here is that a large chain molecule can carry more specificity information than a small one, exactly in proportion to its length. A 50-atom polypeptide chain might carry, say, 3 or 4 amino acid groups and might be one of some 10^4 alternative polypeptides of about the same size; but a 500-atom polypeptide chain, ten times as long, would then be one of about 10^{40} alternatives. The "information" is additive; the specificity is multiplicative.

Informational specificity is of course only a theoretical measure until it is actually expressed in a sequence of functioning chemical groups which control at some time or another the pattern or sequence of some real reactions. This chemical, or mechano-chemical, method of "translation" of the code into real operations is not given by information theory, but is the heart of the biochemical problem; unfortunately it is not yet known for any biochemical code. But it is the biochemical specificity that results that is the operationally significant thing; that is, the specificity of enzymes, or of antibodies, or of inheritance, and so on. We see that a mere 10^4

alternative antibodies from a 50-atom molecular system would not be enough to produce different foreign-tissue reactions, say, to all of the people in a small city. But 10^{40} alternative antibodies, from a 500-atom polypeptide, would be more than enough to give a different chemical label for every cell in every organism on earth.

9. *Self-replication*

This is a consequence of 8, for a sufficiently long sequence, thanks to the famous theorem of von Neumann (1951). This theorem says that in a "suitable environment" of component parts, a sufficiently complex operating "machine" or "automaton" with a "sufficiently long"—but finite!—set of instructions can duplicate itself, including the set of instructions. This is just a mathematical "existence theorem" and does not specify how long the sequence of instructions must be, for a given complexity of machine and for a given level of pre-organization of the component parts themselves. But it does assure us that self-replication is a feat possible in systems of large chain molecules beyond some minimum critical size, where it is not possible for smaller molecules in the same environment.

Quite possibly this minimum critical size—the threshold for autocatalysis—might be rather small for a chain molecule in a soup of amino acids in sunlight, and might be considerably larger—perhaps the size of a phage particle—in the competition and urgency of a bacterial medium. But whatever it is in a given environment, it is clear that this threshold marks an essentially discontinuous change in properties, and that fully-complex molecules larger than this size differ from all smaller ones in a property of central importance for biology.

3. Other Cooperative Properties

10. *Long-range order*

Long-range simple order can be a property of crystals of small molecules. Large molecules, either chains or networks that have repeating units, can also show long-range simple order in their secondary and tertiary folds and helices, as seen from their X-ray patterns; but, in addition, their crystals and complexes can show a long-range complex order in the matching up of specific sequences. The presence of the backbone structures obviously reduces the number of degrees of freedom and simplifies the establishment of the secondary or tertiary or higher-order structures.

The form of the long-range order in a large molecule—whether it is organized into a regular helix rather than into a random coil, for example—is evidently related to the state of contraction of the system, so that it may

be solvent-dependent and time-dependent in a biological system. If the form is variable, of course the properties related to it are variable also; and it is now fairly clear that the presence or absence of secondary and tertiary helices affects such properties as solubility and resistance to chemical attack, optical rotation and hypochromism of the spectra (Kasha & Rich, private communication; Tinoco, 1960 and in press), and the state of aggregation and so the spectra—or "metachromasy"—of any adsorbed dyes (Bradley, in press). These changes of properties are presently interpreted by the helix-random-coil picture (Kasha & Rich, private communication; Tinoco, 1960 and in press); but the possible appearance of such cooperative properties should not depend much on the topology of the system and might be observed in pleated-sheet networks also.

The thing of principal importance for our discussion, however, is the realization that small chains, of less than about 10 side-groups, whether polypeptides or polynucleotides, do not seem to form the secondary and tertiary structures that larger chains can form, and so do not show this dramatic variability of chemical and optical and other properties.

Long-range complex order is also a fascinating property of the very largest molecules. A dramatic example is that of collagen chains, which have a complex sequence with a 2,800 Å repeat unit (Hodge & Schmitt, 1960). It is not clear what this long sequence "means" biologically, but it is clear that it controls parallel "packing" or "crystallization" of the molecules into collagen bundles with 700 or 1,400 or 2,800 Å longitudinal units to give a long-range complex order on a scale that no array of smaller molecules would achieve.

11. *Polyelectrolyte behavior*

There is an extensive literature on properties of polyelectrolytes that go beyond the properties of their monoelectrolyte groups. These include the dependence of their molecular dimensions and rigidity on counterelectrolyte concentration; and the possible chromatographic or molecule-sorting properties of polyelectrolytes like those of the ion-exchange resins. Another important property they might have is the property of forming self-micelles. The monoelectrolyte salts of the fatty acids, sulfonates, and so on, with hydrocarbon chains of C_{10} or longer, form quasi-spherical aggregates, or micelles, in water. The hydrocarbon tails point inward and the polar groups form a hydrophilic shell on the outside. The interior of the micelle is capable of dissolving large quantities of hydrocarbons, fats and oils. It is obviously possible that a long-chain polyelectrolyte containing both hydrophobic and hydrophilic groups, might fold its electrolyte residues similarly back and forth on a two-dimensional quasi-spherical or other surface between hydrophilic and hydrophobic regions, thus stabiliz-

ing a water-lipid interface and perhaps forming a self-micelle. The interior and exterior of a DNA double helix might almost be thought of in this way. Polyelectrolytes do seem to have a number of unusual biological properties, but their detailed configurations are only conjectural in most of these cases.

12. *"Wrap-around" cooperative effects on binding energies and activation barriers*

We come finally to some other properties which large molecules have been conjectured to have and which they may have theoretically, but where the experimental biochemical evidence is conceivably subject to other interpretations and where careful and definitive experiments are still badly needed.

One such theoretically possible property is that of cooperative effects of widely separated chemical groups on a large molecule. This is a familiar concept in discussions of enzyme or antibody action and specificity. Such cooperating groups could bind a substrate molecule at two more points, the binding at one point perhaps facilitating the binding at another if the substrate molecule is a conjugated system, for example. Likewise the cooperative action of such groups could produce a great lowering of activation barriers for certain reactions of the substrate. The linking of the chemical groups by means of the large-molecule backbone would be a way of guaranteeing the connectedness of the two groups or even an exact "wrap-around" spacing. This would greatly reduce the entropy and diffusional factors that might keep similar small-molecule groups from cooperating in a reaction with the same high probability.

Syntheses of long polypeptides of known structure are only just now reaching the point where such long-range cooperative effects in large-molecule reactions might be studied in chemical detail. Large flexible two-dimensional structures of course might show such effects also.

13. *Possible selective transfer of electrons through large ions and radicals*

"Electron mediation", or charge-transfer into one end of a conjugated system and out elsewhere, seems to have been demonstrated in the catalytic action of certain ring systems on small-molecule reactions (Fraser, Sebera & Taube, 1959), and has been conjectured for the shifts of spectra of carotenoids (Platt, 1959). It should not take place very easily in a non-conjugated system, either along saturated bonds or along unsaturated but tightly-bound bonds such as those of the peptide group.

However, it can in principle take place to great distances through ion or radical molecules that have already gained or lost an electron. In such odd-electron systems, the wave function of the highest electron is not localized

but extends throughout the whole molecule and can be modified everywhere by a perturbation at any point. The motion of such an electron (or "hole") to an attracting site in a field is very similar to the motion of an ordinary electron (or hole) through solids or liquids; except that the motion might tend to be "along" the long chain rather than transverse or random through the medium; and the initial and final localization of the electron might have the same kind of specificity that other long-chain reactions can have.

Electron-transfer to or through radicals is now widely supposed to be important in biological processes. The main difficulty in invoking it, so far, is that, as a general principle by itself, it does not seem to lead to any new knowledge about specific reaction intermediates or reaction paths, and so has not led to any hypothetical conclusions that are very definitively testable.

4. Biological Puzzles and Physical Mechanisms

Probably a number of other special properties of large molecules can be added to the list given here. Perhaps sometime it may also be useful to distinguish more closely between the properties of molecules in different size ranges, in the 500-atom range, the 5,000-atom range, the 50,000-atom range, and so on. A different kind of study would be that of the natural and necessary changes with size of the properties of aggregates of different fully complex molecules, such as aggregates of DNA, RNA and protein. But the present list already makes it clear that several biological phenomena which have seemed somewhat peculiar are simply the natural and necessary consequences of increases in single molecule size and complexity.

In particular, the list seems to point the way toward clearing up the third of the three biological puzzles that Szent-Gyorgyi has pointed out (private communication). These puzzles are: (i) how concentration gradients are set up; (ii) how electrical potential differences are produced; and (iii) how muscle moves. The first two of these puzzles have been explained by Burton, (1959) who shows them to be necessary properties of a flowing and chemically reacting system—an open system—in the steady state. Concentration differences and potential differences would not be present, of course, in any closed system in equilibrium. Their presence in biology is puzzling only because we tend to assume that a Donnan equilibrium is applicable, when it is actually not applicable to any open or steady-state system.

This explanation enables us to see that the third of Szent-Gyorgyi's puzzles, the puzzle of muscle action, is of a different type. It appears that properties 1, 3, 5 and 10 of Table 2 are closely related to this problem and may be central in any description of the physics and chemistry of the operation of muscle. Our surprise at the behavior of muscle is then not

due to an unconscious error in our assumptions, as in applying closed-system rules to an open system, but rather to our failure so far to develop as a systematic discipline the study of the mechano-chemistry of long chains with chemically reactant side groups.

Evidently many of the strange phenomena of biology actually follow quite naturally from physics and chemistry, once we have adjusted our concepts to the changed conditions and the changed scale and the changed complexity. What is most important is that this makes it less necessary for us to be dependent on vitalist or epigenetic or other peculiarly biological explanations that have been proposed for these phenomena.

I am indebted to Professor Albert Szent-Gyorgyi for raising the question of the "unique" properties of large molecules, and to Professor Michael Kasha for beginning to answer it for small polyatomics, which was what suggested that the same kind of analysis could be carried one stage further. I also wish to thank my other colleagues at the Summer 1960 Conference at the Institute for Muscle Research at Woods Hole, Massachusetts, for a number of helpful discussions of these questions. Preparation and publication was assisted by a contract with the U.S. Atomic Energy Commission.

REFERENCES

Ames, B. N., Garry, B. & Herzenberg, L. A. (1960). *J. Gen. Microbiol.* **22,** 369.
Bertalanffy, L. von (1952). "Problems of Life", Chapter 5. Harper's, New York.
Burton, A. (1959). *J. Gen. Comp. Physiol.* **14,** 256.
Demerec, M. & Hartman, P. E. (1959). *Ann. Revs. Microbiol.* **13,** 377.
Fraser, R. T. M., Sebera, D. K. & Taube, H. (1959). *J. Amer. chem. Soc.* **81,** 2906, 5514.
Hodge, A. J. & Schmitt, F. O. (1960). *Proc. nat. Acad. Sci. Wash.* **46,** 186.
Kolos, W., Roothaan, C. C. J. & Sack, R. A. (1960). *Rev. mod. Phys.* **32,** 178.
Koshland, Jr., D. E. (1958). *Proc. nat. Acad. Sci. Wash.* **44,** 98.
von Neumann, J. (1951). In "Cerebral Mechanisms and Behavior". John Wiley, New York.
Platt, J. R. (1955). *Proc. nat. Acad. Sci. Wash.* **41,** 181.
Platt, J. R. (1956). *Amer. Scient.* **44,** 180.
Platt, J. R. (1959). *Science,* **128,** 372.
Platt, J. R. (1960). *Bull. Lowell Obs.* **4,** 278.
Platt, J. R. (1961). In "Handbuch der Physik", Part 3, Vol. XXXVII, Chapter 2. Springer, Berlin.
Rosen, R. (1959). *Bull. math. Biophys.* **21,** 387.
Schrödinger, E. (1956). "What is Life?" Doubleday, Garden City, New York.
Thompson, D'A. W. (1959). In "On Growth and Form", Chapter 2. 2nd Edition Cambridge University Press, London. (Also reprinted in J. R. Newman, "The World. of Mathematics", Vol. II, p. 1001. Simon and Schuster, New York.)
Tinoco Jr., I. (1960). *J. chem. Phys.* **33,** 1332.
Tinoco, Jr., I. *J. Amer. chem. Soc.* In press.

J. Theoret. Biol. (1961) 1, 359–381

Tracer Dynamics: II. The Limiting Properties of the Tracer System

Per-Erik E. Bergner

Department of Medical Physics, Karolinska Institutet, Stockholm, Sweden

(*Received* 5 *December* 1960)

This continuation of a previous paper is mainly concerned with an analysis of the "finite tracer system", i.e. the system composed of a finite number of mutually exclusive and uniform compartments (here called subsystems). Special attention is paid to the limiting properties of that system and to the connection between the tracer and its mother substance. The results of the analysis are summarized in the form of a number of theorems and it is shown that their validity can be extrapolated to the more general tracer system with non-uniform compartments. The consequences of that are briefly indicated: it is shown that the possibility of treating non-uniform compartments seems to imply a simplification of the theory in a case like the closed system of plasma and red blood cells. A fundamental property of the present approach is that the fluxes are expressed as functions of the total amounts and not of the concentrations of the actual substance, and, as pointed out in the introduction, this indicates an improvement as regards the significance of the observed variables.

1. Introduction

This communication is a continuation of a previous one (Bergner, 1961), hereafter referred to as I, and familiarity with the methods and symbols given there is assumed.

In I the concept of a general compartment C_ν was introduced as a set of points in the four dimensional state space Q of u, where u is an individual atom with fixed atomic number. The set of all such u in C_ν is equal to $\mathbf{B}_\nu = B_\nu^0 \cup B_\nu$, where B_ν^0 = the set of all naturally present u-isotopes in C_ν (the *mother substance*) and B_ν = the set of those u in C_ν belonging to that u-isotope which is used as *tracer*; the two sets are disjoint, i.e. $B_\nu^0 \cap B_\nu = \phi$, ϕ = the null set. As a result of the approach outlined in I it was found that, on considering the fluxes of u into and out from C_ν, the fundamental variables are the contents of $\mathbf{B}_\nu$ and $\mathbf{B}'_\nu$; here $\mathbf{B}'_\nu$ is the set of u corresponding to the complement C'_ν of C_ν relative to S, i.e. $C'_\nu = S - C_\nu$, with S now standing for the set of points in Q corresponding to the whole system.

This, however, is completely opposite to the view usually adopted: in thermodynamics as well as in common kinetics (statistical or phenomenological) the fundamental variables are the concentrations of the actual species and not their total amounts. However, in biological systems, a macroscopic volume† is usually not uniform and the experimentally observed concentration has therefore not the same significance as that given by the theory. On the other hand, the total amount of a certain kind of particle is usually well defined for any volume (i.e. compartment), not only theoretically but also experimentally.

Let in S a certain set of atoms $\mathbf{B}_s$ be selected, i.e. let a certain atomic number be fixed. This gives a definition of B_s^0, and thus also of B_ν^0 for any compartment $C_\nu \subset S$. Hence, we have a special "population" of particles, namely that of all $u \in B_\nu^0$. It was shown in I that, if the two postulates given there are accepted and if S satisfies certain restrictive conditions relative to B_s^0‡, then there exists a *turnover factor* λ_ν^0 which to a certain extent describes the "dynamic state" of that population. In some respects λ_ν^0 recalls a concept like temperature: λ_ν^0 is an average of a property possessed by each $u \epsilon B_\nu^0$, i.e. the property of having a probability of leaving C_ν during a short time interval.

The significance of λ_ν^0 is thus independent of any assumed models, and the problem is how to determine its numerical value. This is the problem that was stated at the end of I, and it will be recalled that in theorem 2§ two sufficient conditions were given for having $\lambda_\nu = \lambda_\nu^0$, where λ_ν is the *tracer turnover factor* of $\mathbf{B}_\nu$ as defined by (35) of I. The problem could be stated more precisely as: (i) when is S such that λ_ν exists, i.e. when is S a *limiting* S; (ii) when are the sufficient conditions in theorem 2 satisfied; (iii) what kind of assumptions and observations are necessary to make possible a numerical estimate of λ_ν.

In an attempt to get an answer to the above questions we shall in the three following sections analyse some of the most important properties of (s), i.e. of the *finite tracer system* (I, page 127). In I (s) was defined as a tracer system in which the number of possible states of $u \epsilon \mathbf{B}_{(s)}$ is finite, which to a considerable extent simplifies the formal analysis; as will be shown in a later section, the validity of some theorems concerning the properties of (s) can be extrapolated to the more general S.

† With a macroscopic volume we here understand a volume that is large enough to render a quantitative observation possible. In thermodynamics the term macroscopic volume usually refers to any volume "containing a number of molecules sufficiently large for microscopic fluctuations to be negligible" (Prigogine, 1955).

‡ In general S must be perfect, i.e. the amount of mother substance in each point of S must be invariant with respect to time (I, page 137).

§ The enumeration of the theorems in this paper is continued from I: thus Theorems 1 and 2 are given in I.

It will be recalled that the physical counterpart of (s) is a tracer system consisting of a finite number of mutually exclusive *subsystems* $\{ (\mathrm{ss})_i \}_{i=1}^{i=r}$ (cf. page 128 of I), or, in a common terminology, uniform compartments; an (ss) could be considered as a set of points in Q, but where all points $\bar{q} \epsilon (\mathrm{ss})_i$ correspond to one and the same state of $u \epsilon \mathbf{B}_{(\mathrm{s})}$. Such systems have been discussed by several authors (for instance Sheppard & Householder, 1951; Rescigno, 1954, 1956; Berman & Schoenfeld, 1956; Hart, 1955, 1957, 1958) and, therefore, some of the theorems which will be formulated in the following sections are not essentially new. However, several of the earlier investigations have been devoted only to special types of (s) or have been given a form that does not satisfy the present purposes.

Finally, the present treatment, as in I, is always based on the assumption that in every point $\bar{q}$ of Q, the content density of tracer is much less than that of the mother substance (cf. (9) of I):

$$\sigma(\bar{q},t) \ll \sigma^0(\bar{q},t) \Big|_{\substack{\bar{q} \epsilon s \\ t \geqslant 0}} \tag{1a}$$

or for (s), with $b_i(t)$ and $b_i^0(t)$ standing for the content of B_i and B_i^0 respectively, i.e. the total amounts of tracer and mother substance in $(\mathrm{ss})_i$,

$$b_i(t) \ll b_i^0(t) \Big|_{\substack{(\mathrm{ss})_i \subset (\mathrm{s}) \\ t \geqslant 0}} \tag{1b}$$

and we consider S and (s) for non-negative t-values only. Further, when discussing (s) it will always be assumed that the subsets $\{ B_i^0 \}_{i=1}^{i=r}$ of $B_{(\mathrm{s})}^0$ are all non-empty, i.e. when $r =$ the number of (ss) $\subset$ (s), then the determinant

$$|\mathbf{b}^0| \equiv \begin{vmatrix} b_1^0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & b_2^0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & b_r^0 \end{vmatrix} \neq 0 \tag{1c}$$

Hence, $\mathbf{b}^0$ is non-singular, an assumption which is partly motivated by (1b), and it has been formulated in order to avoid any "trivial" forms of (s).

2. (s) After the Supply of Tracer has Ceased

As an evident generalization of (24) of I we have, for $r =$ the number of (ss) $\subset$ (s)

$$\begin{aligned} \dot{b}_i &= \sum_{j=1}^{r} \zeta_{ij} + \zeta_{ie} \\ \dot{b}_i^0 &= \sum_{j=1}^{r} \zeta_{ij}^0 + \zeta_{ie}^0 \end{aligned} \tag{2}\dagger$$

† A dot over the symbol indicates differentiation with respect to t.

where ζ_{ie} and ζ^0_{ie} are non-negative and equal to the fluxes of tracer and mother substance respectively from the environment e of (s) into $(ss)_i$ (cf. page 125 of I). ζ_{ij} and ζ^0_{ij} are the corresponding fluxes from $(ss)_j$ to $(ss)_i$ as defined by (18) of I†. It will also be recalled that

$$\begin{matrix}\zeta_{ij}\\ \zeta^0_{ij}\end{matrix} \geqslant 0, \; i \neq j \qquad \begin{matrix}\zeta_{ii}\\ \zeta^0_{ii}\end{matrix} \leqslant 0 \tag{3}$$

where $|\zeta_{ii}|$ and $|\zeta^0_{ii}|$ are the fluxes out from $(ss)_i$ (cf. page 132 of I).

According to (18) of I, (2) can be written as

$$b_i = \sum_{j=1}^{r} \zeta^0_{ij} A_j + \zeta_{ie} = \sum_{j=1}^{r} \lambda_{ij} b_j + \zeta_{ie} \tag{4a}$$

$$b^0_i = \sum_{j=1}^{r} \lambda_{ij} b^0_j + \zeta^0_{ie} \tag{4b}$$

$$(i = 1, 2, \ldots, r)$$

where λ_{ij} is the *flux constant*‡ for the flux from $(ss)_j$ to $(ss)_i$ and A_i is the so called specific acitivity b_i/b^0_i. In matrix notation (4) gives

$$\dot{\bar{b}} = \xi^0 \cdot \bar{A} + \bar{\zeta}_{\cdot e} = \boldsymbol{\lambda} \cdot \bar{b} + \bar{\zeta}_{\cdot e} \tag{5}$$

for the tracer, and

$$\dot{\bar{b}}^0 = \boldsymbol{\lambda} \cdot \bar{b}^0 + \bar{\zeta}^0_{\cdot e} \tag{6}$$

for the mother substance. Equation (5) is similar to (22) of I derived by stochastic procedure, but here (5) has been given a more general form by adding the column vector $\bar{\zeta}_{\cdot e} = \begin{pmatrix}\zeta_{1e}\\ \vdots\\ \zeta_{re}\end{pmatrix}$. Otherwise the notations are analogous to those used in I, e.g. $\bar{b}$ is a column vector of order r, $\begin{pmatrix}b_1\\ \vdots\\ b_r\end{pmatrix}$, while $\boldsymbol{\lambda}$ is a square matrix of order $r \times r$,

$$\begin{pmatrix}\lambda_{11} & \lambda_{12} & \ldots & \lambda_{1r}\\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2r}\\ \cdot & \cdot & \cdot & \cdot\\ \lambda_{r1} & \lambda_{r2} & \ldots & \lambda_{rr}\end{pmatrix}$$

† When no risk of confusion seems to exist we shall in the following write ζ_{ij} for $\zeta_{ij}(t)$ and b_i for $b_i(t)$ and so on for other similar functions of t.

‡ In I λ_{ij} was called the "probability factor" for the flux $(ss)_j \rightarrow (ss)_i$. However, as the present treatment is purely deterministic, this term seems to be out of place.

The components of $\bar{A}$ are the specific activities, and this vector is connected to $\bar{b}$ through the simple transformation

$$\bar{b} = \mathbf{b}^0 \cdot \bar{A}, \quad \mathbf{b}^0 \equiv \begin{pmatrix} b_1^0 \, 0 \, \ldots\ldots 0 \\ 0 \, b_2^0 \, \ldots\ldots 0 \\ \ldots\ldots\ldots \\ 0 \, 0 \ldots\ldots b_r^0 \end{pmatrix} \tag{7a}$$†

Hence, from (5)

$$\boldsymbol{\lambda} \cdot \mathbf{b}^0 = \xi^0 \tag{7b}$$

Throughout this section we shall assume $\bar{\zeta}_{\cdot e}|_{t \geqslant 0} = 0$, i.e. no tracer is given to (s) after $t = 0$. Hence (5) gives

$$\dot{\bar{b}} = \boldsymbol{\lambda} \cdot \bar{b} \tag{8}$$

As (s) is always assumed to be *stationary* (I, page 127), i.e. $\dot{\boldsymbol{\lambda}}|_{t \geqslant 0} \equiv 0$, equation (8) is a system of r linear and homogeneous differential equations of first order with constant coefficients, which immediately gives:—

THEOREM 3

Let r be the number of (ss) of which (s) consists. If $\zeta_{\cdot e}|_{t \geqslant 0} = 0$ and if $b_i(0) =$ the content of B_i at $t = 0$, then

$$b_i(t) \equiv \sum_{l=1}^{l-n \leqslant r} \beta_{il} e^{-\alpha_l t}|_{t \geqslant 0} \tag{9}$$

with $0 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_n$ and $\beta_{il} = \sum_{j=1}^{r} b_j(0) \eta_{ijl}$, where $\eta_{ijl} = M_{ijl}(t) \cos u_l t + N_{ijl}(t) \sin u_l t$ with $u_l \neq 0$ only if $\alpha_l > 0$. Here u_l and α_l are algebraic functions of $\boldsymbol{\lambda}$, and $M_{ijl}(t)$ and $N_{ijl}(t)$ are polynomials in t of degree $h_l - 1$ ($\sum_{l=1}^{n} h_l = r$), the coefficients of which are algebraic functions of $\boldsymbol{\lambda}$.

Proof

The principles of the proof are well known and we shall therefore not work it out in detail but merely point out those portions which are of special interest to the present case (see also the proof of theorem 6). Thus the characteristic equation of (8) is

$$|\boldsymbol{\lambda} - x\mathbf{J}| \equiv \begin{vmatrix} (\lambda_{11} - x) & \lambda_{12} & \ldots \lambda_{1r} \\ \lambda_{21} & (\lambda_{22} - x) & \ldots \lambda_{2r} \\ \ldots & \ldots & \ldots \\ \lambda_{r1} & \lambda_{r2} & \ldots (\lambda_{rr} - x) \end{vmatrix} = 0 \tag{10}$$

† According to (1c) $\mathbf{b}^0$ is non-singular.

where J stands for the unit matrix (Aitken, 1954). From the general theory of differential equations (see for instance Agnew, 1942) it follows that $\{ -\alpha_l \}$ are equal to the real parts of the roots of (10) and that $\{ h_l \}$ are the multiplicities of the corresponding roots. Hence, what theorem 3 states is that the roots of (10) are all of the form

$$x_l = -\alpha_l + i.u_l \tag{11}$$

where $\alpha_l \geqslant 0$, $i = \sqrt{-1}$ and $u_l \neq 0$ only if $\alpha_l > 0$; thus, β_{il} contains trigonometric terms only if $\alpha_l > 0$. This is a property of $\bar{b}(t)$, the solution of (8), which does not follow from pure mathematical considerations but is a consequence of the physical background of (8), and which therefore has to be taken into account.

Now, as $\bar{\zeta}_{.l}|_{t \geqslant 0} = 0$, the general law of conservation of mass implies†

$$\sum_{i=1}^{r} b_i = b_{(s)}\Big|_{t \geqslant 0} \leqslant 0 \tag{12}$$

This is a necessary condition for (s) being a real physical system, and, together with (4), it gives

$$\sum_{i=1}^{r} \sum_{j=1}^{r} \zeta_{ij}^0 A_j = \sum_{i=1}^{r} \sum_{j=1}^{r} \lambda_{ij} b_j \Big|_{t \geqslant 0} \leqslant 0$$

Hence

$$\sum_{j=1}^{r} [\zeta_{jj}^0 + \sum_{\substack{i=1 \\ i \neq j}}^{r} \zeta_{ij}^0] A_j = \sum_{j=1}^{r} [\lambda_{jj} + \sum_{\substack{i=1 \\ i \neq j}}^{r} \lambda_{ij}] b_j \Big|_{t \geqslant 0} \leqslant 0 \tag{13}$$

Lemma 1

The elements of λ must satisfy the relationship

$$\lambda_{jj} \leqslant -\sum_{\substack{i=1 \\ i \neq j}}^{r} \lambda_{ij}, \; j = 1, 2, \ldots, r \tag{14}$$

and analogous to ξ^0. When (s) is closed (14) is valid as an equality for all j-values.

Proof. We assume (s) to be a realizable physical system, i.e. (13) is valid, and it is also evident that (14) is *sufficient* for (13). On the other hand (13) must hold for every $t \geqslant 0$, i.e. also for $t = 0$. Further, we have put no restrictions on the initial condition $\bar{b}(0)$, except that (1)

† $\sum_{i=1}^{r} b_i = b_{(s)}(t) \equiv b_{(s)}$ is the total amount of tracer present in (s) at time t, i.e. the content of $B_{(s)}$.

must be satisfied. So on choosing $\bar{b}(0)$ such that $b_j(0) = 0$ for all $j \neq k$, (13) gives

$$[\lambda_{kk} + \sum_{\substack{i=1 \\ i \neq k}}^{r} \lambda_{ik}]b_k\,(0) \leqslant 0$$

But this is true for any $k = 1, 2, \ldots, r$; hence (14) is also *necessary*, as (s) is assumed to be a realizable physical system. That a condition analogous to (14) holds for ζ^0 as well, is evident from (13).

When (14) holds as an equality for all values of j, it follows from (12) and (13) that no tracer can leave (s), i.e. (s) is *closed*. This is a mere definition of a closed (s) and it is in agreement with the definition given in I (page 129): (s) is closed if and only if it contains all the states in which $u \epsilon \mathbf{B}_{(s)}$ can be found.

Now, it is evident that we can write λ_{ii} in the form

$$-\lambda_{ii} = \sum_{\substack{j=1 \\ j \neq i}}^{r} \lambda_{ji} + \lambda_{ei} \tag{15}$$

where λ_{ei} is the flux constant for the flux from $(ss)_i$ to e. Hence, after multiplying both members of (15) by b_i or b_i^0 (18) of I gives

$$-\zeta_{ii} = \sum_{\substack{j=1 \\ j \neq i}}^{r} \zeta_{ji} + \zeta_{ei} \tag{16}$$

$$-\zeta_{ii}^0 = \sum_{\substack{j=1 \\ j \neq i}}^{r} \zeta_{ji}^0 + \zeta_{ei}^0 \tag{17}$$

What lemma 1 then states is merely that λ_{ei}, ζ_{ei} and ζ_{ei}^0 are all non-negative and equal to zero for all values of i when (s) is closed, i.e. (s) is closed when $\bar{\lambda}_e. = \bar{\zeta}_e. = \bar{\zeta}_e^0. = 0$†. But, it is not impossible that even if (s) is closed in the sense used here, there are influxes and effluxes of other species than u. Therefore, this definition of a closed (s) is not in agreement with that usually accepted in thermodynamics (see for instance Prigogine, 1955): a system is closed if it has exchange of energy but not matter with its environment. Now, the definition given by lemma 1 has some formal insufficiencies, thus it does not lead to any effective definition of an open (s), a problem to which we shall return in the next section.

The reason why lemma 1 is of interest is not so much that it gives a formal definition of the closed (s), but that we need it in order to

† $\bar{\lambda}_e. \equiv \begin{pmatrix} \lambda_{e1} \\ | \\ \lambda_{er} \end{pmatrix}$, and analogous to the other vectors.

conclude the proof of theorem 3. It can be shown (see Brauer, 1946) that the roots of (10) are situated on or within one of the circles

$$| z - \lambda_{jj} | \leqslant \sum_{\substack{l=1 \\ l \neq j}}^{r} \lambda_{lj}, \; j = 1, 2, \ldots, r \tag{18}$$

where z is a variable in the complex plane. Together with lemma 1 this theorem then immediately gives lemma 2.

Lemma 2

The roots of (10) cannot be pure imaginary and their real parts are non-positive.

This completes the proof of theorem 3. That (14) is sufficient has formerly been pointed out by Berman & Schoenfeld (1956). That (14) is necessary is a consequence, incidentally, of the fact that no restrictions are given in theorem 3 about the way in which $\bar{b}(0)$, the initial condition, is to be chosen, i.e. the tracer may be added to any (ss) at will; an analogous assumption will be formulated and used in the next section. In fact, (14) is identical with (16) of I, which was derived, as a necessary condition, from the conservation of probability. Hence, the fundamental property of probability being non-negative and never greater than unity thus gives the same condition on λ as does the law of conservation of mass.

3. Open and Closed (s)

As was pointed out in the previous section the definition of a closed (s) is not satisfactory and another approach seems indicated. We shall begin with a formulation of two assumptions which seem reasonable from an intuitive point of view: what kind of fundamental properties of (s) can the experimentalist be expected to assume *a priori* (and which will form the basis for his design of experiment). When such assumptions have been given the problem will be to find the formal properties of (s) which are necessary to render the assumptions valid.

Let $\bar{\zeta}_{\cdot e}$ in (5) be such that $\dot{\bar{\zeta}}_{\cdot e} \,|_{t \geqslant 0} = 0$ and with its elements small enough to preserve the validity of (1).

Assumption 1

For every stationary (s) there exists always a uniquely defined point $\bar{b} = \bar{b}_{\infty}$ where $\dot{\bar{b}} = 0$. When (s) is open this is true for all $\bar{\zeta}_{\cdot e}$ such that (1) is valid and $\dot{\bar{\zeta}}_{\cdot e} \,|_{t \geqslant 0} \equiv 0$.†

† Evidently $\dot{\bar{b}} = 0$ is an asymptotic state, and the assumption implies that (s) does not contain any irreversibly absorbing (ss), i.e. states which $u \epsilon \mathbf{B}_{(s)}$ can enter but not leave. When (s) is closed, this is also a consequence of (1c).

That (s) must be open when $\bar{\zeta}_{\cdot e} \mid_{t \geqslant 0} \neq 0$ and $\dot{\bar{b}} = 0$ is self-evident, as the tracer then must be able to leave (s). So on putting $\dot{\bar{b}} = 0$, (5) takes the form

$$\lambda \cdot \bar{b} = -\bar{\zeta}_{\cdot e} \tag{19}$$

If $\bar{\zeta}_{\cdot e} \neq 0$ this system has a unique and non-trivial solution $\bar{b} = \bar{b}_{\infty}$ when λ has the rank r. On the other hand, if the rank is less than r, this would necessarily imply that some of the equations of (19) were linearly dependent. But this is not possible, as, according to assumption 1, all the $\{ \zeta_{ie} \}$ are independent of each other, i.e. within certain limits they can be chosen at will. Therefore, when (s) is open λ must have the rank r.

Let (s) be closed. Then $\dot{\bar{b}} = 0$ is a possible state only if $\bar{\zeta}_{\cdot e} \mid_{t \geqslant 0} = 0$, i.e. after the supply of tracer has stopped, and (19) takes the form

$$\lambda \cdot \bar{b} = 0 \tag{20}$$

Further

$$\sum_{i=1}^{r} b_i = b_{(s)}(0) \tag{21}$$

where $b_{(s)}(0)$ the total amount of tracer in (s) at $t = 0$. If $b_{(s)}(0) > 0$, (20) is possible only if the rank of λ is less than r.

Assumption 2

To any λ_{ii} of a closed (s) it is possible to add $-\lambda_{ei} < 0$, i.e. (s) may be changed to an open (s); the distribution of $u \epsilon B^0_{(s)}$ is kept unaltered by a "compensating supply" of mother substance to $(ss)_i$.

When $-\lambda_{ei}$ is added to λ_{ii}, ζ^0_{ei} changes from zero to a positive value (cf. (17)). This of course means a disturbance in the initial distribution of $u \epsilon B^0_{(s)}$, and what has been assumed is that this disturbance can be eliminated by arranging a supply ζ^0_{ie} equal to ζ^0_{ei}.†

Let (s) be closed and λ having the rank $r - 2$, and all (ss) are kept unchanged but one, $(ss)_j$ say, which is made open, i.e. $-\lambda_{ej}$ is added to λ_{jj}. The rank of the new λ will be at most $r - 1$. But (s) is now open and its λ must have the rank r, hence, assumption 2 implies that the rank of λ must be $r - 1$ when (s) is closed, and so we can formulate

THEOREM 4

(s) is open only if λ and ζ^0 have the rank r, where r is the number of (ss) of (s).

(s) is closed only if λ and ζ^0 have the rank $r - 1$.

† This is necessary, as a change in the distribution may cause a change in the dynamic state of (s) (cf. I, pages 135 and 138).

That λ and ξ^0 have the same rank follows from (7b) as the rank of a matrix is invariant under multiplication by a non-singular matrix (Aitken, 1954).

This theorem is a consequence of the two assumptions; it is in fact equivalent to them in the respect that, as shown in the next section, it implies the existence of the state $\dot{\bar{b}} = 0$. The two assumptions constitute a certain kind of "operational" definition of the concepts "open" and "closed" (s). The definition is in some respects comparatively restrictive. This is especially evident when the present definition of a closed (s) is compared with that in the previous section. Let for instance (s) be closed according to lemma 1, and let the system consist of two parts, $(s)_1$ and $(s)_2$, isolated from each other in such a way that a $u \epsilon \mathbf{B}_{(s)1}$ can never enter into $(s)_2$ and vice versa. Hence, if $(s)_1$ is made open according to assumption 2, then (s) will be neither open nor closed in the sense of assumption 1. However, such a restrictive definition as the one introduced in this section seems necessary, in order to prevent the inclusion of trivial cases which would make the analysis unnecessarily complex.

According to the definition on page 137 of I, we define: (s) is said to be *perfect* when it is stationary and $\dot{\bar{b}}^0 = 0$, i.e. both λ and ξ^0 are independent of t. Thus, when (s) is perfect, (6) takes the form

$$\lambda \cdot \bar{b}^0 = - \bar{\zeta}^0_{\cdot e} \tag{22}$$

For an open (s), $\bar{\zeta}^0_{\cdot e} \mid_{t \geqslant 0} \neq 0$,† while $\bar{\zeta}^0_{\cdot e} \mid_{t \geqslant 0} = 0$ when (s) is closed and perfect, and for any perfect (s), open or closed,

$$\sum_{i=1}^{r} b_i^0 = b^0_{(s)} = \text{constant} > 0 \tag{23}$$

Hence, if (s) is open and perfect, (22) gives

$$b_i^0 = \frac{-1}{|\lambda|} \sum_{l=1}^{r} \zeta^0_{le} \, |\lambda_{li}| \tag{24a}$$

and from (19)

$$b_i = \frac{-1}{|\lambda|} \sum_{l=1}^{r} \zeta_{le} \, |\lambda_{li}| \tag{24b}$$

where $|\lambda_{li}|$ is the co-factor corresponding to the element λ_{li} of the determinant $|\lambda|$. So, when (s) is perfect and $\bar{\zeta}_{\cdot e} \mid_{t>0} \neq 0$,

$$\frac{b_i}{b_i^0} = \frac{\sum_{l=1}^{r} \zeta_{le} \, |\lambda_{li}|}{\sum_{l=1}^{r} \zeta^0_{le} \, |\lambda_{li}|} \tag{25}$$

§ This is necessary as, according to (1c), $|\mathbf{b}^0| \neq 0$.

Now, it seems to be generally held that in a perfect (s), to which the tracer is given continuously, the specific activities $\{ A_i \}$ approach the same value in all $(\mathrm{ss})_i \subset (\mathrm{s})$, i.e. for large values of t one would have

$$\frac{A_i}{A_j} = 1 \text{ for all } \frac{(\mathrm{ss})_i}{(\mathrm{ss})_j} \subset (\mathrm{s}) \tag{26}$$

However, from (25) it follows that (26) is generally valid under special conditions only.

THEOREM 5

For an open and perfect (s) it is generally true that

$$\lim_{t \to \infty} \frac{A_i}{A_j} = \frac{A_{i\infty}}{A_{j\infty}} = 1 \quad \text{for all } \frac{(\mathrm{ss})_i}{(\mathrm{ss})_j} \subset (\mathrm{s}) \tag{27}$$

if and only if

$$\bar{\zeta}_{\cdot e} = R \cdot \bar{\zeta}^0_{\cdot e} \,|_{R>0} \tag{28}$$

where $R = A_{i\infty}$ $(i = 1, 2, \ldots, r)$ is a constant which, in order to preserve the validity of (1b), usually has to be much less than unity. Also, when (28) is satisfied, $\mathbf{b}$ is necessarily non-singular.†

Thus, in order to use (27) in the discussions of experimental data, the tracer has to be distributed to (s) in a particular way, i.e. in a manner identical with that of the mother substance. For example, it would not be sufficient to give an animal tracer by frequent intravenous injections while the mother substance is given via the ordinary food. Experiments, using the technique of frequent administration of tracer, have been reported for instance by Hevesy (1948), and theorem 5 is similar to the conclusions of Berman & Schoenfeld (1956).

Finally we observe the simple but important fact that if (s) is *perfect and closed*,

$$\sum_{j=1}^{r} \zeta^0_{ij} = \sum_{j=1}^{r} \zeta^0_{ji} = 0 \tag{29}$$

This follows from (2), which gives

$$b^0_i = \zeta^0_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{r} \zeta^0_{ij} + \zeta^0_{ie}, \; i = 1, 2, \ldots, r \tag{29'}$$

Then because (s) is perfect and closed, $b^0_i = \zeta^0_{ie} = 0$; whence (29′) reduces to $\zeta^0_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{r} \zeta^0_{ij}$. Comparing this with (14) (the equality condition), then gives the desired result.

† $\mathbf{b} = \begin{pmatrix} b_1 & 0 & \ldots & 0 \\ 0 & b_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & b_r \end{pmatrix}$ is non-singular as a consequence of (27) (1c).

4. The Limiting State of (s)

In the preceding section we formulated two assumptions, one of them was that a unique state $\dot{\bar{b}} = 0$ exists under certain conditions for every stationary (s), and which led to theorem 4. The purpose of the present section is to give a further analysis of some of the properties which are characteristic of the state $\dot{\bar{b}} = 0$.

Now, the general solution of (5) can be written in the form (Bellman, 1960)

$$\bar{b}(t) = e^{\lambda \cdot t} . \bar{b}(0) + \int_0^t e^{\lambda \cdot (t-s)} \, \bar{\zeta}_{\cdot e}(t-s) \, ds \tag{30}$$

where $e^{\lambda \cdot t} = \mathbf{J} + \lambda . t + \frac{1}{2!} \lambda^2 . t^2 + \ldots .$, $\mathbf{J}$ = unit matrix. When $\bar{\zeta}_{\cdot e}|_{t \geqslant 0} = 0$, the solution $\bar{b}(t)$ satisfies (8), hence, from theorem 3,

$$e^{\lambda \cdot t} . \bar{b}(0) = \left\{ \sum_{l=1}^{n} \eta_{ijl}(t) e^{-\alpha_l t} \right\}_{j=1}^{i=r} . \bar{b}(0)$$

where $\left\{ \quad \right\}_{j=1}^{i=r}$ is a matrix of order $r \times r$, and on carrying out the multiplication with the column vector $\bar{b}(0)$, we get

$$e^{\lambda \cdot t} . \bar{b}(0) = \left\{ \sum_{j=1}^{r} b_j(0) \sum_{l=1}^{n} \eta_{ijl}(t) . e^{-\alpha_l \cdot t} \right\}_{i=1}^{i=r} \tag{31}$$

where $\{ \eta_{ijl}(t) \}$ may be polynomials in t and contain sines or cosines functions of t. Thus, by analogy with the preceding, when $\dot{\bar{\zeta}}_{\cdot e} \mid_{t \geqslant 0} \equiv 0$, (30) can be written as

$$b_i(t) = \sum_{j=1}^{r} b_j(0) \sum_{l=1}^{n} \eta_{ijl}(t) e^{-\alpha_l \cdot t} + \sum_{j=1}^{r} \zeta_{je} \sum_{l=1}^{r} \int_0^t \eta_{ijl}(t-s) . e^{-\alpha_l \cdot (t-s)} ds \tag{32}$$

$$(i = 1, 2, \ldots, r)$$

When $\bar{\zeta}_{\cdot e} \mid_{t \geqslant 0} = 0$, $\lim_{t \to \infty} b_i(t)$ is determined by the first double sum only. This limit always exists as, according to theorem 4, (10) can at most have one zero root, i.e. if $\alpha_1 = 0$ then all $\{ \eta_{ij1} \}$ are independent of t. On the other hand theorem 4 states that $\alpha_1 > 0$ when (s) is open, and in that case $\lim_{t \to \infty} b_i(t)$ is dependent on the second double sum only. This double sum contains integrals, which can be of two fundamental types

$$\int_0^t (t-s)^n . e^{-\alpha_l \cdot (t-s)} \, ds \Big|_{n \geqslant 0} \qquad \int_0^t (t-s)^n \sin [u_l . (t-s)] e^{-\alpha_l \cdot (t-s)} ds \Big|_{n \geqslant 0}$$

both of them being convergent for $t \to \infty$. Thus, as a consequence of theorem 4, $\bar{b}(t)$ always has a finite and unique limit, and we can summarize in the form of the following theorem.

THEOREM 6

Let (s) be stationary and open or closed according to theorem 4. Let further $\dot{\bar{\zeta}}^0_{\cdot e} \mid_{t \geq 0} = 0$ and $\bar{\zeta}_{\cdot e} = R \cdot \bar{\zeta}^0_{\cdot e}$ according to (28) but where $R > 0$ only if (s) is open (when (s) is closed no tracer is supplied after $t = 0$). Then,

(*a*) $\lim_{t \to \infty} b_i(t)$ always exists for all $(ss)_i \subset (s)$ and is:

(i) greater than zero and proportional to $b_{(s)}(0)$ when (s) is closed and $b_{(s)}(0) > 0$;

(ii) equal to zero when (s) is open and $\bar{\zeta}_{\cdot e} \mid_{t \geq 0} = 0$;

(iii) greater than zero and independent of $\bar{b}_{(s)}(0)$ but a linear function of $\bar{\zeta}_{\cdot e}$ when (s) is open and $\bar{\zeta}_{\cdot e} \mid_{t \geq 0} \neq 0$.

(*b*) $\lim_{t \to \infty} (b_i / b_j)$ exists greater than zero and independent of $\bar{b}_{(s)}(0)$ for all $\frac{(ss)_i}{(ss)_j} \subset (s)$ when (s) is closed and $b_{(s)}(0) > 0$, or when (s) is open and $\bar{\zeta}_{\cdot e} \mid_{t \geq 0} = 0$. In the latter case the limit is equal to a quotient between two linear functions of $\bar{\zeta}^0_{\cdot e}$.

That $\lim_{t \to \infty} b_i(t)$ is proportional to $b_{(s)}(0)$ when (s) is closed does not follow from the above analysis, but is a consequence of the fact that, in this case, $\lim_{t \to \infty} \bar{b}(t)$ is a solution of (20) and (21). That $\lim_{t \to \infty} b_i(t) > 0$ when (s) is closed and $b_{(s)}(0) > 0$ is a consequence of (1c) and of theorem 7 below.

It should be observed that the open (s) to which no tracer is given after $t = 0$ is not included in (*b*). For instance, $a_1 > 0$ may correspond to a complex root of (10), and for such an (s) we will then have $\lim_{t \to \infty} (b_i / b_j)$ in the form

$$\frac{M_i \sin u_1 t + N_i \cos u_1 t}{M_j \sin u_1 t + N_j \cos u_1 t}$$

which in general lacks any definite limiting value. On the other hand, when $R > 0$, the diagonal matrix $\mathbf{b}$ is non-singular according to theorem 5, hence $\lim_{t = \infty} b_i > 0$, and so $0 < \lim_{t = \infty} (b_i / b_j) < \infty$.

Assume (s) being perfect and $\bar{\zeta}_{\cdot e} \mid_{t \geq 0} = 0$. Then, from (5) we get for $\dot{\bar{b}} = 0$,

$$\xi^0 . \bar{A} = 0 \qquad (33)$$

According to theorem 4 this equation has only a trivial solution $\bar{A} = o$ when (s) is open, while when (s) is closed there exist a non-trivial solution $\bar{A} = \bar{A}_\infty$, which is uniquely determined by (33) and

$$\sum_{i=1}^{r} b_i = \sum_{i=1}^{r} b_i^0 . A_i = b_{(s)}(o) \tag{34}$$

This agrees with theorem 6. From (29) it follows that the sum of all elements in each column and each row of ξ^0 is equal to zero, i.e.

$$\xi^0 . \{ 1 \} = \{ 1 \}^T . \xi^0 = o \tag{35}$$

where $\{ 1 \}$ is a column vector of order r with all its elements equal to unity, and $\{ 1 \}^T$ = the transpose of $\{ 1 \}$. Thus $\bar{A}_\infty = h . \{ 1 \}$, with the scalar h determined by (34). This means that when $b_{(s)}(o) > o$, the only possible $\bar{A}$, for which $\dot{\bar{b}} = o$, corresponds to a state such that the specific activities are the same in all subsystems. So, at $\dot{\bar{b}} = o$, we have (cf. (25) of I)

$$\mathcal{J}_{ij} = A_\infty . (\zeta_{ij}^0 - \zeta_{ji}^0) \tag{36}$$

where $\mathcal{J}_{ij} = \zeta_{ij} - \zeta_{ji}$, i.e. the net flux of tracer from $(ss)_j$ to $(ss)_i$. It is thus evident that $\dot{\bar{b}} = o$ does not necessarily imply any "detailed equilibrium" in the sense that all the net fluxes $\{ \mathcal{J}_{ij} \}$ should vanish.

The term "equilibrium" for the state $\dot{\bar{b}} = o$ should therefore be avoided, otherwise there is a risk of confusion with the same term in thermodynamics. Only the population of $u \epsilon \mathbf{B}_{(s)}$, and no other species, is here considered; hence we may have a flux of matter and energy into and out from (s) even when (s) is closed with respect to $u \epsilon \mathbf{B}_{(s)}$ (see page 365). Therefore, when $\dot{\bar{b}} = o$ it is possible in principle to have a "cyclic equilibrium" of the type shown by Fig. 1, with a corresponding production of

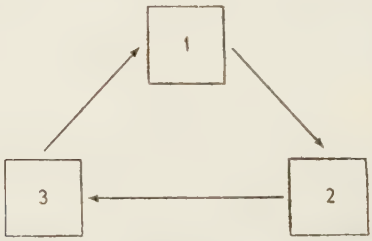

FIG. 1.

entropy. Such a state does not correspond to what is usually meant by an "equilibrium" (Denbigh, 1951), thus the *steady state* seems as a more appropriate name for $\dot{\bar{b}} = o$.

THEOREM 7

When (s) is perfect and closed the steady state condition $\dot{\bar{b}} = 0$ is satisfied if and only if

$$\frac{A_i}{A_j} = 1 \text{ for all } \begin{matrix}(ss)_i \\ (ss)_j\end{matrix} \subset (s) \tag{37}$$

Only when (s) is *symmetric*, i.e. ζ^0 is symmetric, this state corresponds to a detailed equilibrium.

It should be observed that theorem 7 does not make impossible unsymmetric and closed (s); it only states that such a system cannot be in detailed equilibrium. A simple example of the latter type of system is given in Fig. 2: a tube bent into a plane vertical rectangle, completely filled

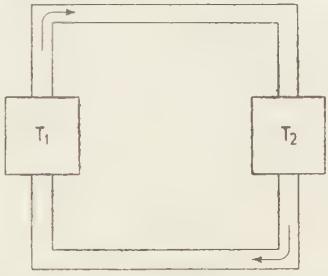

FIG. 2.

with a liquid or gas solution of a substance B^0 and its tracer B, and where the temperature $T_1 > T_2$ and the arrows indicate the directions of the net flows of **B**. This system can be in a steady state but not in detailed equilibrium.

As an example of a symmetric (s) we have the ion exchange system treated by Harrison *et al.* and Boyd *et al.* (1959), where (s) is closed; the simple perfect mammillary system, with a flow between e and (s) only at its central subsystem, is a good example of an open and symmetric (s). Such systems, i.e. perfect and symmetric, have been discussed, from a different point of view in an interesting paper by Hearon (1953) where it is shown that equality in (14) together with the condition of detailed equilibrium is enough to guarantee that the roots of (10) are all real. Now, as shown by Hearon, this is true also for an (s) when its λ can be written in the form $\lambda = \lambda^* - \lambda_{\cdot e}$, where λ^* corresponds to a symmetric, perfect and closed (s) and $\lambda_{\cdot e}$ is a diagonal matrix of non-negative elements. So, for this special class of tracer systems, to which the mammillary system belongs, the parameters $\{\beta_{il}\}$ of (9) will not contain any periodic functions of t.

5. The Limiting Properties of S

Following the methods developed in the last section of I we write (cf. (34) of I)

$$\varphi_{ji}(t) \equiv \int_{(ss)_i} f_i(\bar{q}, t)d\bar{q} \int_{(ss)_j} d'\lambda(\bar{q}', \bar{q}) = \lambda_{ji} \mid_{i \neq j} \tag{38}$$

$d'\lambda(\bar{q}', \bar{q}).\delta t$ is equal to the probability of a transition from $\bar{q}$ to $\bar{q}'$ during the short time interval δt, and the integration over $(ss)_j$ is performed with respect to $\bar{q}'$. It will be recalled that $f_i(\bar{q}, t) \equiv \dfrac{\sigma(\bar{q}, t)}{b_i(t)}$, and thus $\int_{(ss)_i} f_i(\bar{q}, t)d\bar{q} \equiv 1$ (cf. (7) and (8) of I). Therefore, if λ_{ji} is independent of the form of $f_i(\bar{q}, t)$, it is necessary and sufficient that $d'\lambda(\bar{q}', \bar{q})$ is independent of $\bar{q} \epsilon (ss)_i$.

It was found in I that (s) is equivalent to a finite Markov chain with continuous time parameters, where then the states are the subsystems. Now, that $d'\lambda(\bar{q}', \bar{q})$ is independent of $\bar{q} \epsilon (ss)_i$ is in fact necessary and sufficient for $(ss)_i$ being a Markovian state. Hence, from the discussion of (38): as the set $\{ \lambda_{ij} \}$ of (s) is independent of the form of the functions $\{ f_i(\bar{q}, t) \}$, (s) is a Markov chain.

For two disjoint compartments C_ν and C_μ of S we get

$$\varphi_{\mu\nu}(t) \equiv \int_{C_\nu} f_\nu(\bar{q}, t)d\bar{q} \int_{C_\mu} d'\lambda(\bar{q}', \bar{q}) \tag{39}$$

The existence of such integrals has always been taken for granted. Hence, it is possible to divide C_ν and C_μ into a finite number of mutually exclusive subsets $\{ C_{\nu k} \}_{k=1}^{k=n_\nu}$ and $\{ C_{\mu l} \}_{l=1}^{l=n_\mu}$ and write (39) in the form

$$\varphi_{\mu\nu}(t) = \sum_{k=1}^{n_\nu} [f_\nu(\bar{q}_k, t)d\bar{q} \sum_{l=1}^{n_\mu} d\lambda(\bar{q}_l, \bar{q}_k)] + \epsilon \tag{40}$$

where, by proper subdivision, ϵ always can be made arbitrarily small; $\bar{q}_k$ and $\bar{q}_l$ stand for an arbitrary point in $C_{\nu k}$ and $C_{\mu l}$ respectively. ϵ can thus be disregarded and we get

$$\varphi_{\mu_\nu}(t) = \sum_{k=1}^{n_\nu} [f_\nu(\bar{q}_k, t)d\bar{q}_k \sum_{l=1}^{n_\mu} \lambda_{lk}] \tag{41}$$

where we have written λ_{lk} for $d'\lambda(\bar{q}_l, \bar{q}_k)$.

In (40), ϵ does not only depend on the subdivision but also on the choice of C_ν and C_μ as well as on the form of $\{ f_\nu(\bar{q}_k, t) \}_{k=1}^{k=n_\nu}$ i.e. of $\sigma(\bar{q}, t) \mid_{\bar{q} \epsilon C_\nu}$. The function $\sigma(\bar{q}, t)$ depends on the dynamic state of S (I page 127) and on the supply of tracer to S. But, in simple terms, to get ϵ small is a mere question of making the subdivision "fine" enough. Therefore, without penetrating further this purely mathematical problem, we state, as an evident fact, that we can always find a division of S into a

finite but presumably large number of mutually exclusive subsets $\{ C_l \}_{l-1}^{l=n}$ so that (41) holds for any two disjoint compartments C_ν and C_μ and for any reasonable supply of tracer. Now, to such a subset C_l there corresponds a set of flux constants $\{ \lambda_{kl} \}$, which are defined in the usual sense as

$$\varphi_{kl}(t) \equiv \int_{C_l} f_l(\bar{q}, t) d\bar{q} \int_{C_k} d'\lambda(\bar{q}', \bar{q}) = \lambda_{kl} \Big|_{l \neq k} \tag{42}$$

Here λ_{kl} is independent of $f_l(\bar{q}, t)$, and so C_l has the same fundamental property as has a subsystem; but for the subset this property is only approximately valid.

Thus, as a consequence of the assumed existence of the integrals of (39) we get: to any S it is always possible to find an (s), the behaviour of which to any degree of accuracy agrees with that of S. Hence, the theorems, which in the preceding sections have been derived for (s), can with certain formal modifications be applied to S, and on this basis some of the properties of S that are most important for the present purposes will be summarized below. This will be done in the form of a number of theorems, but, as these are obvious reformulations of the preceding results, the theorems will mostly be stated without formal proof.

Corresponding to (1c) we always assume

$$\sigma^0(\bar{q}, t) \Big|_{\substack{\bar{q} \in s \\ t \geqslant 0}} > 0 \tag{43}$$

If $d\zeta^0(\bar{q}, e) \mid_{\bar{q} \in s}$ stands for the flux of mother substance from e to $\bar{q} \epsilon S$ we get from (28)

$$d\zeta(\bar{q}, e) = R . d\zeta^0(\bar{q}, e) \Big|_{R>0} \tag{44}$$

and it is always assumed that

$$\frac{d}{dt} [d\zeta^0(\bar{q}, e)] \Big|_{\substack{\bar{q} \in s \\ t \geqslant 0}} \equiv 0 \tag{45}$$

The definitions of open and closed (s) in terms of the Assumptions 1 and 2 can be applied to S†, but then $-\lambda_{ei}$ is replaced by $-\lambda(e, \bar{q})$, defined according to (27) of I as

$$\lambda(e, \bar{q}) \equiv \int_e d'\lambda(\bar{q}', \bar{q}) \Big|_{\bar{q} \in s} \tag{46}$$

with $\bar{q}'$ as the variable of integration. The corresponding fluxes are

$$\begin{aligned} d\zeta^0(e, \bar{q}) &= \lambda(e, \bar{q}) . \sigma^0(\bar{q}, t) d\bar{q} \\ d\zeta(e, \bar{q}) &= \lambda(e, \bar{q}) . \sigma(\bar{q}, t) d\bar{q} \end{aligned} \tag{47}$$

† S does then not contain any absorbing states (cf. footnote on page 366).

If it is also assumed that

$$d\zeta^0(\bar{q}, e)\Big|_{\substack{\bar{q}\,\epsilon\, s \\ t \geqslant 0}} \begin{cases} \equiv 0 \text{ when } S \text{ is closed} \\ > 0 \text{ for some } \bar{q} \text{ when } S \text{ is open} \end{cases} \tag{48}$$

then we have the same conditions put on S as were put on (s) in theorem 6 (cases (i) and (iii)). Hence, with $b_\nu' =$ the content of $B_\nu = B_s - B_\nu$, we get

THEOREM 8

S is stationary, (44), (45) and (48) hold and $b_s(0) > 0$ when S is closed. Then

(*a*) $\lim_{t\to\infty} b_\nu(t)$ exists greater than zero and is:

(i) proportional to $b_s(0)$ when S is closed;

(ii) independent of the initial conditions but a function of $d\zeta(\bar{q}, e)$ when S is open.

(*b*) $\lim_{t\to\infty} \dfrac{b_\nu'}{b_\nu} \equiv \beta_\nu$ where $0 < \beta_\nu < \infty$ exists independent of $b_s(0)$, but is a function of $d\zeta^0(\bar{q}, e)$ when S is open.

We may chose C_ν as small as we please, hence, (*a*) has the following corollary:—

S is stationary, (44), (45) and (48) hold and $b_s(0) > 0$ when S is closed. Then

$$\lim_{t\to\infty} \sigma(\bar{q}, t)\Big|_{\bar{q}\,\epsilon\, s}$$

exists greater than zero and with the same properties as given for $\lim_{t\to\infty} b_\nu(t)$.

Let B_s^* be any subset of $\mathbf{B}_s$ and let $B_s^{*\prime} = \mathbf{B}_s - B_s^*$. If $\sigma^*(\bar{q}, t) \ll \sigma^{*\prime}(\bar{q}, t)\Big|_{\substack{\bar{q}\,\epsilon\, s \\ t \geqslant \infty}}$ the above corollary holds for $\lim_{t\to\infty} \sigma^*(\bar{q}, t)$ as well. But this is possible if and only if $\lim_{t\to\infty} \sigma^0(\bar{q}, t)$ exists greater than zero.

THEOREM 9

(*a*) The conditions (45) and (48), together with that of S being stationary, imply that S approaches a perfect S when $t \to \infty$.

(*b*) When S is perfect it is stationary and (45) and (48) hold.

As we shall in the following mainly concern ourselves with the limiting properties of S, it is thus evident that assuming S to be perfect implies no serious restriction.† Hence, theorems 8 and 9 immediately give rise to the following two theorems.

† In I (page 138) it was stated that the assumption that S is stationary but not perfect seems in fact more restrictive than that S is perfect when being stationary.

THEOREM 10

S is perfect, (44) holds and $b_s(0) > 0$ when S is closed. Then

$$\lim_{t\to\infty} \frac{b_\nu}{b_\nu} = 0$$

THEOREM 11

S is perfect, (44) holds and $b_s(0) > 0$ when S is closed.
Then
(*a*) S is a limiting S, i.e.

$$\lim_{t\to\infty} f_\nu(\bar{q}, t) \equiv \pi_\nu(\bar{q}), \lim_{t\to\infty} f'_\nu(\bar{q}, t) \equiv \pi'_\nu(\bar{q})$$

$$0 < \frac{\pi_\nu(\bar{q})}{\pi'_i(\bar{q})} < \infty$$

Both exist independent of $b_s(0)$, but are functions of $d\zeta^0(\bar{q}, t)$ when S is open.
(*b*)

$$\lim_{t\to\infty} \varphi_\nu(t) \equiv \lambda_\nu, \lim_{t\to\infty} \varphi'_\nu(t) \equiv \lambda'_\nu$$

Both exist and λ_ν and λ'_ν are the tracer turnover factors of $\mathbf{B}_\nu$ and $\mathbf{B}$, respectively.

It will be recalled that $\varphi_\nu(t)$ and $\varphi'_\nu(t)$ are the tracer turnover functions of $\mathbf{B}_\nu$ and $\mathbf{B}'_\nu$ respectively (cf. (28) and (29) of I): $\varphi_\nu(t)\cdot\delta t$ is the average value of the probability of a $u\epsilon B_\nu$ to leave C_ν during the short time interval δt, while $\varphi'_\nu(t)\cdot\delta t$ is the average value of the probability of a $u\epsilon B'$ to enter C_ν during δt, the averages being taken over B_ν and B'_ν respectively at time t.

Now, the condition

$$\frac{A_i}{A_j} = 1 \text{ for all } \frac{(\mathrm{ss})_i}{(\mathrm{ss})_j} \subset (\mathrm{s}) \tag{49}$$

in theorems 5 and 7 can be written for S as

$$\left(\frac{\sigma(\bar{q}, t)}{\sigma^0(\bar{q}, t)} \Big/ \frac{\sigma(\bar{q}', t)}{\sigma^0(\bar{q}', t)}\right)\Bigg|_{\substack{\bar{q}\,\epsilon\, s \\ \bar{q}'\,\epsilon\, s}} = 1$$

or

$$\frac{\sigma(\bar{q}', t)}{\sigma(\bar{q}, t)} = \frac{\sigma^0(\bar{q}', t)}{\sigma^0(\bar{q}, t)}\Bigg|_{\substack{\bar{q}\,\epsilon\, s \\ \bar{q}'\,\epsilon\, s}}$$

Hence

$$\int_C \frac{\sigma(\bar{q}', t)}{\sigma(\bar{q}, t)}\, d\bar{q}' = \int_{C_\nu} \frac{\sigma^0(\bar{q}', t)}{\sigma^0(\bar{q}, t)}\, d\bar{q}'\Bigg|_{\bar{q}\,\epsilon\, s}$$

which means that (49) corresponds to

$$f_\nu(\bar{q}, t) = f_\nu^0(\bar{q}, t)\Big|_{\bar{q}\,\epsilon\,s} \tag{50}$$

(cf. I, eq. (5)). Therefore, theorems 5 and 7 give, according to condition (i) in Theorem 2, the following:

THEOREM 12

If S is a limiting and perfect S, then

$$\lambda_\nu = \lambda_\nu^0,\ \lambda_\nu' = \lambda_\nu^{0\prime}$$

when (44) holds.

From (30) of I we have that for any compartment $C \subset S$ the continuity equation can be written as

$$\dot{b}_\nu = \varphi_\nu' \cdot b_\nu' - \varphi_\nu \cdot b_\nu + \zeta_{\nu e} \tag{51}$$

where $\zeta_{\nu e}$ can be defined as

$$\zeta_{\nu e} \equiv \int_{C_\nu} d\zeta(\bar{q}, e) \tag{52}$$

On dividing both members of (51) by b_ν and letting $t \to \infty$, we get from theorems 8, 10 and 11

$$\lambda_\nu = \beta_\nu \cdot \lambda_\nu' + \chi_\nu \tag{53}$$

with $\chi_\nu \equiv \lim_{t\to\infty} \dfrac{\zeta_{\nu e}}{b_\nu}$.

The sufficient conditions for this equation are given by the corresponding theorems.

If $b_\nu^0 \gg b_\nu^{0\prime}$ in a perfect S, it is not impossible that experimentally we get $\beta_\nu = 0$. On the other hand,

$$\chi_\nu = \lim_{t\to\infty} \frac{\zeta_{\nu e}}{b_\nu'} \cdot \frac{b_\nu'}{b_\nu} = \beta_\nu \lim_{t\to\infty} \frac{\zeta_{\nu e}}{b_\nu'} = 0,\dagger$$

and so (53) would give $\lambda_\nu = 0$. This is probably the situation for the exchange of certain ions between bone minerals and the surrounding fluid (Neuman & Neuman, 1958), and in such a case we could speak of C_ν as a compartment with *apparent absorbing states*.

Finally we observe that, if there is a non-negative and finite number t_0 such that $\dot{\varphi}_\nu\Big|_{t\geqslant t_0} = 0$, eq. (51) gives,

$$\left.\begin{aligned} \dot{b}_\nu &= \varphi_\nu' \cdot b_\nu' - \lambda_\nu \cdot b_\nu + \zeta_{\nu e} \\ \ddot{b}_\nu &= \dot{\varphi}_\nu' \cdot b_\nu' + \varphi_\nu' \cdot \dot{b}_\nu' - \lambda_\nu \cdot \dot{b}_\nu \end{aligned}\right|_{t\geqslant t_0} \tag{54}$$

† According to theorem 8, $\lim_{t\to\infty} b_\nu' > 0$.

as $\dot{\zeta}_{\nu e} \equiv 0$. Therefore, if $\varphi'_\nu(t)$ is known for some value of $t \geqslant t_0$, and $b_\nu(t)$ and $b'_\nu(t)$ are known (from experimental data), then $\varphi'_\nu(t)$ appears as the solution of a first order differential equation (in (54) λ_ν can be eliminated).

6. Concluding Remarks

In the preceding section it was stated that to any S it is possible to find an (s), the behaviour of which agrees with that of S. Thus, the real system could, as a principle, be thought of as being composed of a finite but presumably large number of subsystems. However, these subsystems have no physical significance of their own, but are merely mental concepts introduced for explaining the behaviour of the system: the stated relationship between S and (s) concerns only the behaviour and not the "structure" of the two systems. That there are systems where the different (ss) can be identified, at least as a good approximation, was indicated in I (page 133), but this does not invalidate the above statement.

(s) is a Markov chain, the states of which are the subsystems $(ss)_1$, $(ss)_2$, ..., $(ss)_r$. Let now the subsystems be lumped together into mutually exclusive classes C_1, C_2, ..., C_n, $n < r$. A system having $\{ C_l \}_{l=1}^{l=n}$ as the set of observable states will in general not be a Markov chain: if the system is in the state C_l, then the probability of a transition, during a fixed time interval, from C_l to another class of (ss) will, in general, depend on the position within C_l, hence C_l is not a Markovian state (cf. page 374). It seems therefore natural to conclude that simple "compartment models" of complex metabolic systems cannot in general be expected to satisfy a linear differential equation of the form given by (5). Also, from this point of view it would follow that the commonly accepted procedure of expressing the observed values of $b_\nu(t)$ in compartment C_ν as $b_\nu(t) = \sum_{l=1}^{n} \chi_l e^{-\alpha_l t}$ ($\{ \dot{\chi}_l \} \equiv 0$, $0 \leqslant \alpha, < \ldots < \alpha_n$) assuming $n =$ the number of compartments of the observed system, has no evident justification. In fact, a compartment can be chosen at will and depends on the experimental procedure, not on the system; as is the case with the subsystems, a compartment therefore lacks any physical significance of its own. But our treatment has here been non-approximative and the conclusions must be regarded in light of this fact. It is evident that the experimental precision is of fundamental importance when discussing experimental data, a problem which will be considered in another paper.

Though being to a large extent tentative the present investigation has given results which seem reasonable and, at least in certain special cases, also of some practical value. As an illustrative example we may take the closed *in vitro* system of plasma and red blood cells, with the plasma as

one compartment, C_p, and the population of cells as the other, C_c. We consider the exchange of ions between C_p and C_c, and (53) takes here the simple form

$$\lambda_p = \beta_p . \lambda_c \tag{55}$$

If $b_s \mid_{t<0} = 0$ and the tracer is supplied to C_p at $t = 0$ in a single dose and in such a way that instantaneous mixing can be assumed, then $\dot{\varphi}_p \mid_{t \geqslant 0} = 0$, and λ_p could be determined by plotting $(\dot{b}_p/b_p)$ versus t and extrapolating to $t = 0$. The system is assumed to be perfect, hence $\beta_p = (b_c^0/b_p^0)$ is easily determined and thus λ_c is given by (55). This also gives the value of λ_c^0, as, according to theorem 12, we have here $\lambda_c = \lambda_c^0$.

In an excellent analysis by Sheppard & Householder (1951) this system has been considered as an (s). These authors discuss the possibility of considering the population of cells as one single (ss), and they show that if the variation of the dynamic properties between the cells is of reasonable size, this is indeed a good approximation. But here each cell is considered as one (ss) only, in spite of the fact that the cells are far from being homogeneous and that their volumes are finite, a problem which in part has been discussed by Harris & Prankerd (1957).

However, in making a general quantitative description, the present approach will to a large extent eliminate such problems as those indicated above. Thus, the significance of λ_c^0 is independent of what the population of cells is like, as far as the system can be considered perfect. For instance, the population may be a mixture of young and old cells or of normal and sick ones, and improvements of the experimental technique will not make necessary any change of model, but merely imply a better estimate of the turnover factors.

Of some general interest are systems where $\dot{\varphi}_\nu$ or $\dot{\varphi}'_\nu$ vanish. One example is the plasma-red cell system treated above, but in general the situation is more complicated, and perhaps (54) may be of some use in such cases where neither λ_ν nor λ'_ν can be determined separately. In fact, this equation seems to point to a possible method for determining the turnover of metabolites with unknown or non-observable precursors. But in any case, it is always necessary to have some *a priori* knowledge of the system and to allow different forms of approximations. Certainly this means a limitation of the applicability of the methods, and the experimental difficulties may be overwhelming. This becomes especially apparent when the experimental consequences of (44) are considered: when the system is open the tracer should be supplied in exactly the same manner as is the mother substance.

In two papers an attempt has now been made to outline a theory which would make it possible to consider metabolic systems without referring

to any uniform compartments. To a large extent the analysis has been conceptual, and it has been based essentially on the two postulates in I and also on some mathematical properties of the transition probabilities, which, however, have been taken for granted (cf. the footnote on page 136 of I). Perhaps some additional postulates would not have been out of place, but, as these would then concern more or less purely mathematical properties only, this point has not been taken into consideration.

I wish to express my thanks to Dr. A. F. Bartholomay, Harvard Medical School, Boston, for advice and encouragement, and to Dr. U. Uhlhorn, Department of Theoretical Physics, Stockholm, for criticism and stimulating discussions. I am also indebted to Civ.Ing. L. Ingelstam of the Royal Institute of Technology, Stockholm, for valuable suggestions concerning some detailed problems of purely mathematical character.

This work has been supported by Contract AF 61 (052)–21, Office of Air Research and Development Command, U.S. Airforce, through its European Office in Brussels, and by the National Institutes of Health, Bethesda Md, Grant D 700.

REFERENCES

Agnew, R. P. (1942). "Differential Equations". McGraw-Hill, New York.
Aitken, A. C. (1954). "Determinants and Matrices". Oliver & Boyd, London.
Bellman, R. (1960). "Introduction to Matrix Analysis". McGraw-Hill, New York.
Bergner, P.-E. E. (1961). *J. theoret. Biol.* **1**, 120.
Berman, M. & Schoenfeld, R. (1956). *J. appl. Phys.* **27**, 1361.
Brauer, A. (1946). *Duke math. J.* **13**, 387.
Denbigh, K. G. (1951). "The Thermodynamics of the Steady State". Methuen, London.
Harris, E. J. & Prankerd, T. A. J. (1957). *J. gen. Physiol.* **41**, 197.
Harrison, E. L., Raymond, W. H. A. & Sutton, A., and Boyd, J., Neuman, W. F. & Hodke, H. C. (1959). On the Mechanism of Skeletal Fixation of Strontium: Parts I and II; *Arch. Biochem. Biophys.* **80**, 97.
Hart, H. E. (1955). *Bull. math. Biophys.* **17**, 87.
Hart, H. E. (1957). *Bull. math. Biophys.* **19**, 61.
Hart, H. E. (1958). *Bull. math. Biophys.* **20**, 281.
Hearon, J. F. (1953). *Bull. math. Biophys.* **15**, 121.
Hevesy, G. (1948). "Radioactive Indicators". Interscience, New York.
Neuman, W. F. & Neuman, M. W. (1958). "The Chemical Dynamics of Bone Minerals". The University of Chicago Press.
Prigogine, I. (1955). "Introduction to Thermodynamics of Irreversible Processes". Charles C. Thomas, Springfield.
Rescigno, A. (1954). *Biochim. biophys. Acta*, **15**, 340.
Rescigno, A. (1956). *Biochim. biophys. Acta*, **21**, 111.
Sheppard, C. W. & Householder, A. S. (1951). *J. appl. Phys.* **22**, 510.

J. Theoret. Biol. (1961) **1**, 382–403

Evolution and the Theory of Games

R. C. Lewontin

Dept. of Biology, University of Rochester, Rochester, New York

(*Received* 6 *March* 1961)

The shortcomings of present population genetic theory are discussed as they pertain to problems of speciation, extinction and the evolution of genetic systems. It is suggested that the modern theory of games may be useful in finding exact answers to problems of evolution not covered by the theory of population genetics. An outline of relevant topics in the theory of games is given. It is suggested that the most pertinent utility measure for a population is its one-generation probability of survival and that a strategy or a mixture of strategies corresponding to a *maximin* strategy will be found in natural populations. These notions are applied to a population segregating for two alleles with different norms of reaction in different environments. For the model chosen the optimal strategy is found to be homozygosis for different alleles in different populations due either to inbreeding or genetic isolation. A segregating polymorphism in such populations would be a detriment to the species, although the heterozygotes are more constant in fitness.

The Present State of Evolutionary Theory

The modern theory of evolutionary dynamics is founded upon the remarkable insights of R. A. Fisher and Sewall Wright and set forth in the *loci classici* "The Genetical Theory of Natural Selection" (1930) and "Evolution in Mendelian Populations" (1931). By the time of the publication of Wright's paper in 1931 all of the theory of population genetics, as it is presently understood, was established. It is a sign of the extraordinary power of these early formulations, that nothing of equal significance has been added to the theory of population genetics in the thirty years that have passed since that time. Yet we cannot take this period to mean that we now have an adequate theory of evolutionary dynamics. On the contrary, the theory of population genetics, as complete as it may be in itself, fails to deal with many problems of primary importance for an understanding of evolution.

The structure of population genetic theory may be briefly summed up as follows. Given an assemblage of organisms defined as a genetic population, given the breeding structure of that population, the frequencies of various alleles, the phenotypes of the various genotypes, the statistical nature of

the environment, the amounts of recombination between loci, the mutation rates and migration rates, then it is possible in theory to predict the genetic structure of the population at some future time. Such a prediction may be unambiguous, as in deterministic theories, or else a statement about the probability that the population will be in a given state, as in stochastic theories. But whether deterministic or stochastic, these theories are restricted to predictions about changes in the genotype composition of a *given population* with *given* forces determining the changes of gene frequency. Population genetics is not genetics *of* populations, but genetics *in* populations. It is the genetics of phyletic change, of the gradual replacement of one set of alleles by another within a phylad genetically continuous in time and space.

A complete theory of evolution, however, must address itself to other problems besides those of phyletic change within populations. Despite the great amount written about speciation, there is as yet no mathematical theory of species formation nor are there rigorous formulations of the process of phyletic extinction. Yet speciation and extinction stand together with phyletic change as the main features of evolution. What theory will account for the origin and rise of individual homeostasis as an evolutionary mode? Have we a theory of the waxings and wanings of sexuality, of the variations in amount of recombination, of rates of mutation, of all those factors which enter as parameters in the formulations of population genetics? Except for intuitive theories, the answer to these questions remains, "No". For example, the genetic conditions which favor species formation have recently been discussed by Mayr (1959), Carson (1959) and Wallace (1959) and each, while arguing cogently, comes to a different conclusion. Another example is the problem of the relative amounts of genetic variability to be expected in marginal populations as opposed to central populations. Dobzhansky (1955) and Carson (1955, 1959) hold essentially opposing views on this question and a third, closer to that of Carson, is given by Lewontin (1957). It is not the purpose of this introduction to examine these questions in detail, but rather to point out the diversity of conclusions arrived at by evolutionists. This diversity arises not from disagreement about facts, but from the lack of a rigorous theoretical framework for solving problems in which characteristics of a population *as a whole* are the basic parameters. Disagreements of this nature do not arise in population genetics, *sensu strictu*, because of the great power of population genetic theory.

Specification of the Problem

The problem that faces the theorist of evolution then, is to account for differences between populations and species in their degrees of

inbreeding and outbreeding, in amount of recombination, in mutation rates, in degree of somatic plasticity as opposed to genetic flexibility, in dispersal rates, in short, in all those characteristics of populations which govern adaptation. The solution of this problem requires two things. First, *some measure of the success or degree of adaptation of a population must be specified.* Thus, it must be possible, in some unambiguous way, to state that a population with free recombination has a success value X, while one with no recombination has a success value Y, when all other conditions are specified. This is analogous to the measure of Darwinian fitness, or adaptive value of a genotype, which is defined in population genetics as the relative probability that an egg of that genotype will produce an egg in the next generation.

Second, given a measure of success or adaptation, *it must be possible to construct an adaptive calculus* which will predict the kinds of genetic structures that populations will evolve under given environmental conditions. In population genetics the adaptive calculus is built upon the basic rules of mendelism and the mathematical techniques are the standard ones of algebra, the calculus and the theory of stochastic processes. When the population as a whole is the unit of enquiry, however, some new basis must be found. Although the results of population genetic theory will play some part, it is possible that the calculus of population evolution will be a radically different mathematical technique, perhaps one that is yet to be invented by mathematicians. It is the purpose of this article to propose one possible mode of attack, provided by the theory of games developed by von Neuman and Morgenstern (1944).

Outlines of a Relevant Game Theory

Game theory is not well known to biologists so that some of its general outlines must be presented here. What follows is not a strictly conventional description of the theory of games and some new notions have been introduced. There are several reasons for this. First, game theory has its roots in the social sciences and therefore much of the terminology carries connotations of purposive behavior. The notions of a "player", "choosing a strategy", "preferring an outcome" are foreign to modern mechanistic biology and so must be either discarded or very carefully redefined. Second, some of the axioms of game theory flow directly from ideas of purposive behavior. The very definition of the important concept of the "utility" of an outcome depends tautologically on human preferences. Third, many of the criteria of optimum solutions to game theoretical problems are inapplicable because they appeal intuitively to human preferences. This last objection is of some help to the evolutionist, because he may reject many of the suggested solutions and narrow down the field of

choice to what may be considered evolutionarily important solutions. Finally, the processes of evolution are sufficiently different from human choice processes, that some new ideas must be introduced into the classical picture of game theory.

In what follows, any divergence from the standard theory of games will be noted when this divergence is significant.

A GAME IN EXTENSIVE FORM

We may imagine one or more players engaged in a game according to a set of rules. Besides the actions of the player or players there are physical conditions, say the order of the cards in a shuffled pack, which determine the events of the game. The game proceeds in discrete steps called *moves* and at each move one or more players performs an *act* A_i from a denumerable set of available acts $\{ A \}$. The model of discrete moves and acts simplifies the description of the game, but a further extension to *time-continuous* games and games in which $\{ A \}$ is a continuous set can and must be made for some biological problems.

In the most general game the set of available acts $\{ A \}$ may be different for each player and moreover may change at each move in a way dependent upon the act performed at a previous move. For example, in bridge the cards available to a player in any given trick are determined by the cards he has played in previous tricks.

At each move the physical conditions extraneous to the players' acts may also change. The set of *states of nature* $\{ N \}$ is formally identical with a set of acts, $\{ A \}$ and in this sense nature may be considered simply another player with its own set of possible acts.

After each move has been completed there may be a result which we shall call a *local outcome* (this term does not exist in standard game theory) which depends upon the Act A_i performed by each player and the state of nature N_i. Whether or not a local outcome does result after every move depends upon the definition of an outcome for each game. In bridge again, the taking of a trick is a local outcome and this will occur only after four moves, rather than one. In addition to local outcomes there will be a *terminal outcome* which may be either a special local outcome signaling the end of the game or else some function of the local outcome which is not itself a local outcome. In bridge the terminal outcome is the total number of tricks taken by a pair of players, or more generally a function of that total and the total contracted for by the bidding. It is important to notice that the local outcome has an intricate relationship with previous and future outcomes and acts. The set of local outcomes $\{ o \}_t$ at any time may depend upon past local outcomes, *even though those past outcomes do not*

effect the act performed by a player at time *t*. Thus, a player may decide to bet half his remaining resources at every toss of a coin. After any given toss, t, the outcome in terms of money lost or gained will be correlated with all previous values of o but these will not in any way have affected the betting pattern of the player. This kind of dependence is of great importance for evolutionary problems. Suppose that population size is the outcome of interest. Population size N_t will depend upon N_{t-1} and the various factors which determine the rate of increase of a population. Yet the genetic structure of a population, which determines its rate of increase, may be totally insensitive to population size.

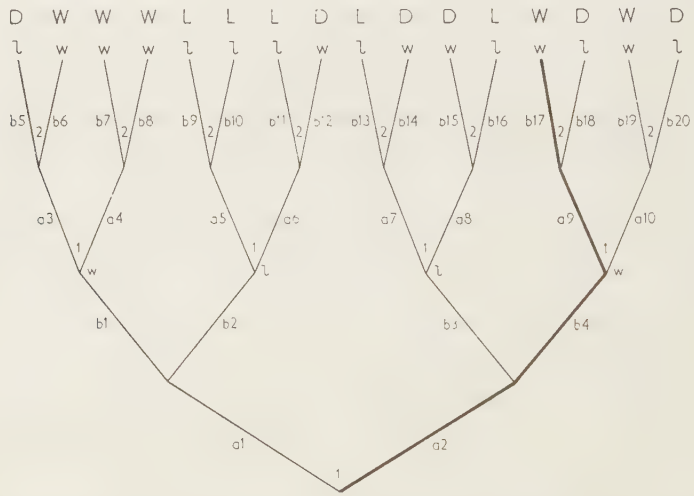

FIG. 1. A game in extensive form. The lower case letters l and w indicate local outcomes after each two moves. The upper case letters indicate terminal outcomes D (draw), W (win) and L (lose). The letters a and b with subscripts are alternative acts for players 1 and 2 respectively. The heavy line is one possible game with player 1 winning.

All of these concepts can be summarized as in Fig. 1, a so-called game tree. It is assumed that there are two players (or one player and states of nature) and that their moves alternate strictly. The number at each branch point is the player number, the letters along each branch are the possible acts available to each player. There is a local outcome after each cycle of two moves and a terminal outcome, win, lose, or draw. Both outcomes are expressed relative to player 1. This figure is not the diagram of a particular set of plays, but of all possible sets given the structure of the game. Any particular play is represented by following one set of branches to the top of the tree. The heavy line shows one such play with player 1 winning. The terminal outcome is defined as the majority of local outcomes.

THE NORMAL FORM OF A GAME

The enumeration of a game in extensive form with all its moves and local outcomes can be reduced to a simpler form, the *normal form*. By tracing any particular path along the game tree a set of sequential acts for a given player can be enumerated. In Fig. 1, for example, the heavy line includes acts a_2 and a_9 for player 1 and b_4, b_{17} for player 2. The set $\{a_2, a_9\}$ is a compound act for player 1 and is called the player's *strategy*. For each player in a game there will be a set of strategies, each strategy corresponding to one of the possible sets of sequential acts. The set of strategies S_i for one player taken together with the set of strategies S_j for the other player will determine a set of terminal outcomes O_{ij}. These can then be put in the form of Table 1, the normal form of a game. Thus, it can be said that if player 1 adopts strategy 1 and player 2 strategy 3 the terminal outcome of the game will be O_{13}.

TABLE 1

The normal form of a game. There are two players with 3 *and* 4 *strategies respectively. The* O_{ij} *are terminal outcomes of a given strategy pair.*

		Player 2			
		S1	S2	S3	S4
	S1	O11	O12	O13	O14
Player 1	S2	O21	O22	O23	O24
	S3	O31	O32	O33	O34

It is not necessary to encompass the entire game in the normal form. Any subset of acts performed in a sequence of moves will form a strategy and to each of these strategy pairs S_i', S_j will correspond an outcome O_{ij}' suitably defined. In the case that there is only a single act comprising a strategy S' the outcome O' will be a *local outcome*, while for larger subsets it will be of the nature of a terminal outcome. That is : O_{ij}' may be some arbitrary function of the local outcomes included in that portion of the game. This notion of a strategy S' encompassing only a subset of sequential acts is not usual in game theory since it is the final outcome of a game which is usually considered. However, for our purposes, it is important to know "who is ahead" at any particular time more often than it is to know "who won". At any particular moment in evolution, populations are in the process of playing the game so that much of our data comes from only partly completed matches. The entire question of the importance of local outcomes and terminal outcomes will be discussed below under the head-

ing of pay-off functions. What is of importance here is that the normal form allows a game extended in time to be represented as a single step game, the analysis of which is much simpler. The complexity of the model has not been lessened by this procedure, however. In place of a series of simple acts and local outcomes we have a smaller set of complex strategies, and terminal outcomes.

The Game Against Nature

In place of one of the players in the two person game of Table 1, we may put the physical universe. In such a model, the *game against nature*, we have a single player adopting some course of action, a strategy, in the face of some state of nature N_j. Depending upon the state of nature, N_j, each strategy of the player will result in a specific outcome O_{ij}. As an example we may take a population of self-fertilized plants which may be homozygous for an allele *a* or *A*. Individual homozygous *aa* produce 5 seeds in dry weather and 10 seeds in wet weather. On the other hand *AA* plants have the reverse norm of reaction. Homozygosity for *A* or for *a* can be considered as alternative strategies for the population. Then the game theoretical form of this situation could be represented in one of the two forms shown in Table 2. Table 2a is the normal form of an extended game of *t* generations duration. In such a case the states of nature are not simply *wet* or *dry* but a complete specification of all the possible sequences of wet and dry seasons. Assuming that the populations in question are growing exponentially, the order of seasons is not relevant, but only the number of wet and dry seasons. There are then $t + 1$ states of nature (0 wet and *t* dry, 1 wet and $t - 1$ dry, ... *t* wet and 0 dry). The terminal outcomes expressed as total population size, are shown in the body of Table 2a.

TABLE 2

Two normal forms of a game against nature. Homozygotes AA *and* aa *have intrinsic rates of increase* 5 *and* 10 *in wet weather and the reverse in dry. Outcomes are expressed as population size. Table 2a shows terminal outcomes after* t *seasons, Table 2b local outcomes after* 1 *season.*

a.

		States of Nature					
		0	1	2 ...	*r* ...	*t*	wet
		t	*t*−1	*t*−2 ...	*t*−*r* ...	0	dry
Strategies	AA	10^t	$10^{t-1}5^1$	$10^{t-2}5^2$	$10^{t-r}5^r$	5^t	
	aa	5^t	$10^1 5^{t-1}$	$10^2 5^{t-2}$	$10^r 5^{t-r}$	10^t	

b.

		States of Nature wet	dry
Strategies	*AA*	5	10
	aa	10	5

Table 2b, gives the local outcomes of a *one-move-game* form of the same situation. There are then only two states of nature, wet and dry, and two strategies. The entire sequence of t generations is then a sequence of games of type 2b. The game shown in 2b is called a *component game* of a *sequentially compounded game.*

SEQUENTIALLY COMPOUNDED GAMES

The sequential compounding of one-move normal games is of very direct interest in the solution of evolutionary problems, so that a description in some detail is worthwhile.

In general, each component game in a sequentially compounded game will have a different set of strategies, states of nature and local outcomes. Even in the simple situation of Table 2b, the outcomes in the second generation are not the same as in the first and more important *depend* upon the *outcome of the first.* Suppose, for example, the population had adopted the strategy *AA* and the weather had been wet in the first generation. Then there would be 5 plants in second generation and the component game for this second generation would have the local outcomes shown in Table 3a. On the other hand, had the weather been dry in the first generation there would be 10 plants in the second generation, so that the second component game would appear as in Table 3b. Usually the dependence will be far more complex than this. In a population segregating for both A and a, there will be changes in gene frequency in successive generations so that the strategy set will also change in a dependent manner.

TABLE 3

Second component games of a sequentially compounded game shown in Table 2b. Table 3a is the second game on the assumption that at the first game strategy AA *was chosen and the weather was wet. Table 3b is the second component game if the weather had been dry in the first season.*

a.

		States of Nature wet	dry
Strategy	*AA*	25	50
	aa	50	25

b.

		States of Nature	
		wet	dry
Strategies	*AA*	50	100
	aa	100	50

In general let Γ^k be the kth alternative component game with a set of strategies $\{ S \}^k$ and states of nature $\{ N \}^k$. The local outcome of a strategy S_i^k and state of nature N_j^k for this kth component game will be represented as

$$o_{ij}^k = a_{ij}^k \text{ and } (P_{ij}^{k0} \Gamma^0, P_{ij}^{k1} \Gamma_{ij}^{k1}, \ldots P_{ij}^{kn} \Gamma^n)$$

The meaning of this rather formidable expression is simple. If component Γ^k is played with strategy S_i^k and state of nature N_j^k actually occurring, then the outcome will be made up of two parts. First there will be an absolute local outcome a_{ij}^k (increase of the population, say). In addition there will be a probability P_{ij}^{k0} of playing component game Γ^0 at the next generation, a probability P_{ij}^{k1} of playing game Γ^1 at the next generation and so on up to a probability P_{ij}^{kn} that the second component game played will be Γ^n. To be completely general we will define Γ^0 as a special *terminal game* consisting of a single outcome and of no further play. Thus P_{ij}^{k0} is the probability that play stops (the population becomes extinct, for example) and that at that moment there is some terminal outcome O. If two species are in competition a terminal game would occur when one of the species became extinct and the terminal outcome with respect to the victor might be a sudden expansion to fill the space formerly preempted by the loser.

There are a number of variations of sequentially compounded games depending upon assumptions about the number of component games, Γ, and the existence of a terminal game, and the existence of absolute local outcomes. These are variously called recursive, stochastic, attrition, and survival games, but the names are misleading in a biological context and they will not be discussed in detail in this introductory essay. Further details can be found in Luce & Raiffa (1957).

The Notion of an Optimal Strategy

Up to this point we have sketched a formalism for describing interactions between populations and states of nature. This formalism must now be put to use for solving some of the problems alluded to in the introduction. For the sake of concreteness, let us consider a clonally reproducing plant occupying an area in which there are both wet and dry localities. In the simplest case, we may consider three genetic strategies

available to this plant: the population consists entirely of *AA* genotypes, entirely of *aa* genotypes, or entirely *Aa* genotypes. Table 4 shows the

TABLE 4

Hypothetical seed set for three genotypes of a clonally reproducing plant in two states of nature.

	N1 wet	N2 dry
AA	15	2
Aa	8	7
aa	4	10

outcomes in terms of seed set per plant in wet and dry localities. In terms used by Dobzhansky & Levene (1955), the homozygotes are "narrow specialists" with a high seed set in one environment and a small seed set in the other, while the heterozygotes are "jacks-of-all-trades" with an intermediate seed set in both environments. We may now ask "which of the three strategies is *best* for the species?" The answer to this question requires two separate decisions. First, what is the relation between outcome values and the values of some success measure of the population? It is necessary to choose a *utility* associated with each outcome, and what is most important, utilities are not necessarily the same as outcomes. It is not true that in all cases a population setting 10 seeds per plant is "better off" than one setting 5 seeds per plant, even if all other things are equal. It is entirely possible that high seed set may result in such overcrowding at germination as to damage the next generation. The problem of choosing a proper numerical utility associated with a given outcome is a biological rather than a mathematical one. We shall discuss this point in a later section. For the moment let us assume that such a relationship between a given outcome and a utility can be made. In the present state of game theory it is necessary to postulate that the *utility function* M defined over a set of alternative outcomes is *linear*. What is meant by a linear utility function is that if outcome O_1 is given utility M_1 and outcome O_2 is assigned utility M_2, then any mixture of these outcomes in proportion p and (1 — p) will have a utility defined as

$$M = pM_1 + (1 - p)M_2$$

Note that the linearity of the utility function is *not* a condition imposed upon the relation of utilities to outcomes. The outcomes 5, 10 and 15 can be assigned the utilities 7, 43 and 12, say, perfectly within the definition of utility. It will certainly be necessary to relax this stricture for some biological problems, but such a relaxation will require certain radical changes in present methods of solution for game theoretical problems.

Having assigned a numerical linear utility function M to the set of alternatives the second decision must be made: What criterion defines the best or *optimal strategy*? In the example of Table 4 assume that the utilities are as in Table 5. Will the population be "better off" if it adopts

TABLE 5

A set of arbitrary utilities illustrating the principle of the admissible set. The admissible set here is (S_1, S_2, S_3). *These three strategies are those of Table 4.*

	N1	N2	*optimality*
S1	8	0	maximax
S2	3	4	maximin
S3	2	7	insufficient reason
S4	2	5	inadmissible
S5	0	3	inadmissible

strategy *AA*, *aa* or *Aa*? If it were sure that the state of nature was "wet" then strategy *AA* would have the highest utility, while *aa* would be best if the state of nature was always dry. On the other hand if the states of nature differed in different localities the answer is not clear. If the probability distribution of the states of nature were known, the problem would be solved by definition of the utility function M. Since the utility function M is linear, a *utility index* can be calculated for each strategy. Let p_j be the probability of the state of nature N_j and let u_{ij} be the utility associated with outcome o_{ij}. Then the utility index of strategy S_i is

$$M(S_i) = p_1u_{i1} + p_2u_{i2} + \ldots + = \sum_{j=1}^{n} p_ju_{ij}$$

That is $M(S_i)$ is the *average utility* of strategy S_i and best strategy, by definition of utility, is the one with the highest value of $M(S_i)$. In Table 5 let the probability of a wet locality be $\frac{1}{4}$ and that of a dry locality $\frac{3}{4}$. Then

$$\begin{aligned} M(AA) &= \tfrac{1}{4}(8) + \tfrac{3}{4}(0) = 2{\cdot}00 \\ M(Aa) &= \tfrac{1}{4}(3) + \tfrac{3}{4}(4) = 3{\cdot}75 \\ M(aa) &= \tfrac{1}{4}(2) + \tfrac{3}{4}(7) = 5{\cdot}75 \end{aligned}$$

and *aa* is the optimal strategy.

In general, then, there is no serious problem of defining an optimum strategy when an *a priori* distribution of states of nature or of the second player's strategies can be specified. The optimal strategy, *by definition of linear utility functions*, is the strategy giving the highest average utility.

Uncertain States of Nature

The real use of game theory becomes clear in situations for which there is no *a priori* distribution of states of nature specifiable. This is certainly

the usual, if not universal, case. The empirical definition of probability depends upon large (virtually infinite) numbers of repetitions of given events. In fact the events of nature are so complex that they are seldom if ever repeated in the lifetime of a population or species. With respect to living organisms nature is *capricious* rather than *random*. What is the probability of a flood, a succession of dry years, of competition with a new invading species of a particular sort? In general such probabilities cannot be specified (or even defined). Since probabilities cannot be specified for different states of nature, no average utility can be calculated so that other criteria of optimality are needed.

(i) *Admissable strategies*

Following the notation of Luce & Raiffa (1957) we will say that for two strategies S and S′

(a) S is *equivalent* to S′($S \sim S'$) if they have identical utilities for every state of nature.

(b) S *strongly dominates* S′($S > S'$) if the utilities of S are greater than for S′ for *every* state of nature.

(c) S *weakly dominates* S′($S \gtrsim S'$) if S has a higher utility than S′ for at least one state of nature and never has a lower utility than S′.

Then the set of *admissable* strategies â* is the set such that *no* strategy S in â weakly dominates any other strategy S′ in â. Whatever criterion of optimality is used, the optimal strategy, must be in the admissible set S*. That is, no reasonable criterion of optimality needs to consider a strategy which is not better than a second strategy in *some* state of nature. Table 5 makes this notion of an admissible set clear. S_5 is first rejected because it is strongly dominated by S_2, S_3 and S_4. It is the worst in every state of nature. S_4 is also to be rejected because it never has a higher utility than S_2 or S_3. It is weakly dominated by S_2 and S_3. The admissible set is then $\{ S_1, S_2, S_3 \}$ because none of these is weakly dominated by any other.

(ii) *The principle of insufficient reason*

It is sometimes stated as a principle of probability theory that if there is no *a priori* reason for assigning specific probabilities to alternative outcomes of a process, they should all be regarded as equally likely. Using such a principle each state of the possible state of nature would be assigned a probability of $1/n$. With this uniform probability distribution an expected utility can be calculated and the optimal strategy determined. The shortcomings of such a criterion are obvious and have been extensively reviewed in the probability literature. One example will suffice to concretize these objections. If two indistinguishable coins are tossed, there will be three

distinguishable outcomes; 2 heads, 2 tails, 1 head and 1 tail. The principle of insufficient reason would result in a probability of $\frac{1}{3}$ for each of these events. If the coins have different dates or one were scratched so that they could now be distinguished, there will be four distinguishable outcomes: head, head; tail, tail; head, tail; tail, head; and each has a probability of $\frac{1}{4}$ by insufficient reason. But this is absurd for either the outcome "two heads" has a probability $\frac{1}{3}$ or a probability $\frac{1}{4}$ and that probability cannot depend upon whether or not the coins are distinguishable.

The counter-intuitive result of the principle of insufficient reason does not, however, rule it out as a method of solution for organisms. It may very well turn out that some species will be found to employ strategies which are optimal under this criterion. If various combinations of temperature, humidity, soil type, other organisms, and so on, really represent distinguishable environments, then over long periods of time only a small fraction of the possible environmental combination will occur and each will occur essentially once. Even lumping environments that are fairly similar as indistinguishable in order to produce a probability distribution will result in a platykurtic function very similar to a uniform distribution. Within the range of extreme environments, then, an *a priori* uniform distribution of states of nature may provide a close approximation to an optimal strategy. In Table 5 strategy S3 is optimal under this criterion.

(iii) *Strategies based on extreme utilities*

If no *a priori* probabilities can be set for the state of nature or for an opponent's choice of strategy, then the complete set of utilities will not really provide any criterion for optimality. In such a case we turn to the extreme values of utilities, the greatest and the least utility that a given strategy could yield. One such criterion is to choose that strategy whose *greatest* possible utility under any alternative state of nature is larger than for any other strategy. This is the *maximax* optimality criterion. In Table 5 strategy S1 is the maximax strategy since it has a utility of 8 under one of the states of nature, which is higher than any utility for any other strategy.

Under what conditions may we expect populations to adopt such a one-sided strategy? One possibility is in a species expanding into an unoccupied territory or a territory in which competitors are relatively weak. In unfavorable environments the invader would have a small rate of increase or perhaps even lose in competition with the incumbent group. In favorable conditions however, there would be an explosive increase in the population and the setting up of many new populations by migration. Even though some of these populations may be extinguished in unfavorable environments, the maximax strategy would assure in the end successful

spread. Such a maximax strategy would be especially important if rapid colonization were important to eventual success since rapid colonization and occupation of the open niche might preclude later successful invasions of other species.

A second optimum strategy based on extreme utilities is one based on a weighting of the maximum and minimum utilities for each strategy. Let α be the weight assigned to the minimum utility of a given strategy m_i and $1 - \alpha$ the weight assigned to the maximum utility of that strategy, M_i. Then an index due to Luce and Raiffe (1957) is

$$H(\alpha) = \alpha m_i + (1 - \alpha)M_i$$

and that strategy is optimal which has the highest $H(\alpha)$. Such an index makes it possible to take into account both favorable and unfavorable states of nature, but the precise criterion for the choice of α is not clear and seems to imply some *a priori* distribution of the best and worst conditions. When $\alpha = 0$ we have the *maximax* criterion discussed above. The most interesting and probably the most useful case for evolutionary problems is the assignment of $\alpha = 1$ so that only the minimum utility is considered. This leads to the *maximin* criterion of optimality.

THE MAXIMIN CRITERION OF OPTIMALITY

When α is set to unity, that strategy is optimum whose lowest utility over all states of nature is larger than for any other strategy. That is, the population is guaranteed a utility m_i at the very least. In Table 5 the maximin strategy is S2. I propose that this maximin strategy is the one which will be found in most cases for reasons which will be discussed below.

For a given matrix of utilities the maximin strategy or strategies can be seen by examination. However, it is possible to increase the *security level* of a game (the maximin utility) by extending the strategies to include so-called *mixed strategies*. In a mixed strategy each of the pure strategy alternatives is assigned a probability p_i and it is assumed that the game is played over and over again with strategy S_i being chosen p_i of the time. The biological analogue of a mixed strategy is a species with many local populations, a proportion p_i of these populations adopting a strategy S_i. The problem is then to find the values of p_i which will maximize the security level of the population. There is a simple method of solution of this problem which *guarantees an average utility to the population at least as high as the maximin pure strategy, irrespective of the distribution of states of nature.* This procedure, because it applies irrespective of the probability of various states of nature, is a very powerful one.

We shall demonstrate the method of solution graphically, although

from the graphical solution it is obvious that what is really involved is the solution of a series of simultaneous linear equations. Figure 2 shows such a solution. The abscissa represents proportions of two strategies S_i and S_k from ·00 S_i, 1·00 S_k to 1·00 S_i, 0 S_k. On the ordinate are utility values. Choosing a particular pair of strategies S_i, S_k, the utilities of pure S_i and pure S_k for a given state of nature N_j will be represented by a point on the left margin and right margin of the diagram respectively. Because utility is defined linearly, the utility of any mixture of S_i and S_k will be the value on the ordinate of the straight line joining the values of the two pure strategy utilities m_i and m_k. A similar line is drawn for every state of nature with respect to these two strategies. In Fig. 2 the two dashed lines

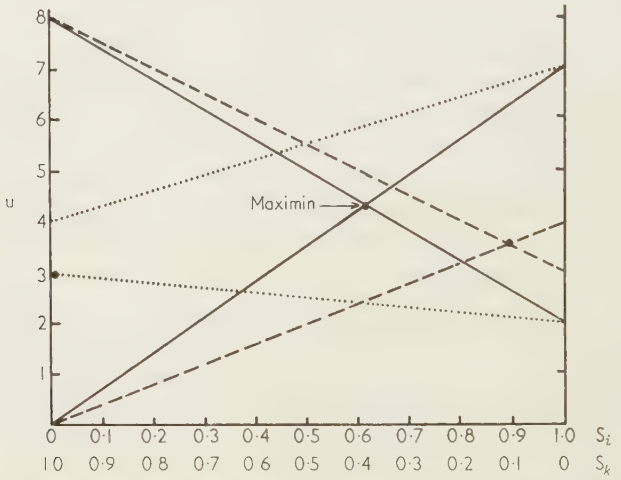

FIG. 2. Graphical solution of the maximin problem for strategies S1, S2 and S3 of Table 5. The abscissa is in varying proportions of S_i and S_k. The ordinate is in utility units. Points of intersection represent the maximin utility and the appropriate weighting of S_i and S_k for a given strategy pair S_i, S_k. Dashed lines are for the pair S2, S1, dotted lines for the pair S3, S2 and solid lines for pair S1, S3.

are those for the two states of nature N1 and N2 in Table 5 with respect to strategies S_2 and S_1. The point of intersection of the lines provides the maximin mixed strategy (·11 S1, ·89 S2) and the security level of this mixture, 3·56. A similar set of lines can be drawn for every pair of strategies S_i, S_k and a maximin strategy for that combination found in the same way. The pair of solid lines in Fig. 2 is for strategies S3 and S1 while the dotted lines are for strategies S3 and S2. Note that the S3, S2 lines do not intersect so that the maximin strategy for this pair is pure S2 with a security level of 3. To determine the maximin mixture of all three strategies the results of the pairs can now be compared. The mixed strategy (·385 S1, ·615 S3) gives the best security level, 4·31 of any pair. Moreover, any

weight given to S2 will lower this security level since substitution of S2 for either S1 or S3 always results in a lower security level *even though* S2 *is the pure maximin strategy*. Thus, no weight should be given this strategy in the optimal strategy mixture and the optimal mixed strategy remains (·385 S1, ·615 S3).

In the case that there are more than two states of nature, there will be several mixed utility lines for each strategy pair, as shown in Fig. 3. The

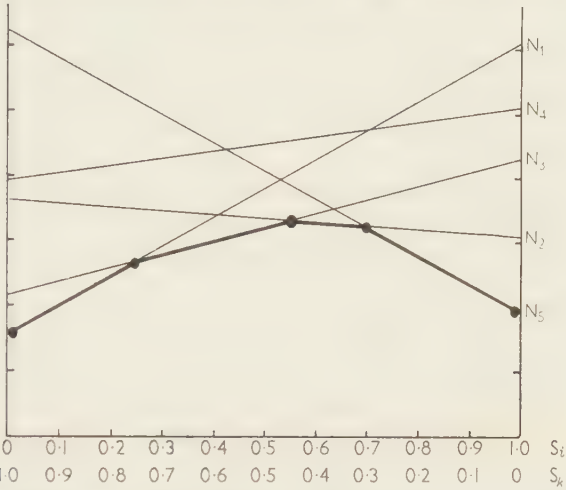

FIG. 3. An example of the linear graphic solution of a maximin problem with more than two states of nature. The dark line is the admissible boundary of solution with the maximin mixed strategy indicated by the large dot.

maximin strategy mixture is that marked by the heavy dot. The heavy line marks off the *admissible boundary* of solutions and the heavy dot is that point on the admissible boundary which has the highest security level.

An Example of an Evolutionary Problem

The artificial example of Table 5 presents an excellent opportunity to demonstrate the application of game theory to an evolutionary problem. It is fairly generally accepted that in diploid sexually-reproducing organisms, homozygotes are more specialized in their adaptive properties than heterozygotes. A heterozygote will have about equal fitness in a great variety of environments, while a homozygote will have a low fitness in most environments but a much higher fitness in some environment to which it is specifically adapted. Consider Table 6 showing the utility in two states of nature of a homozygote *AA*, a heterozygote *Aa*, and the other homozygote *aa*. The utilities are taken from Table 5 and we have already demonstrated that although the heterozygote is the maximin *pure* strategy,

a greater security level can be obtained by a mixed strategy in which about 40% of the population of the species are homozygous *AA* and 60% homozygous *aa*. A population which was heterozygous *Aa* completely (clonal reproduction) would lower the security level of the entire species. But we may carry the analysis a step further. Nothing has been said about a polymorphic population in which *AA*, *Aa*, and *aa* are all present in a segregating mixture. It is quite possible that polymorphism for each population would be a better strategy than the mixed one suggested above. It is important to note that a segregating population is adopting a *pure* strategy, *not a mixed* one in our sense. If segregating populations are allowed, then there are a great number of added pure strategies S4. . . . S*n* each representing some set of proportions of *AA*, *Aa* and *aa* in a population.

To solve this problem we will assume that there is no facilitation among genotypes. That is, if there is a proportion X_1 of *AA*, X2 of *Aa* and X3 of *aa* in a population, the utility of this polymorphic population is

$$m = X_1 m_{AA} + X_2 m_{Aa} + X_3 m_{aa}$$

for any state of nature. Thus for states of nature N_1 and N_2 respectively:

$$m_1 = 2X_1 + 3X_2 + 8X_3$$
$$m_2 = 7X_1 + 4X_2 + 0X_3$$

The values of X_1, X_2 and X_3 will be determined by the gene frequency of *A*, say *q*, and the degree of inbreeding *F*. In general:

$$X_1 = q^2(1 - F) + Fq$$
$$X_2 = 2q(1 - q)(1 - F)$$
$$X_3 = (1 - q)^2(1 - F) + F(1 - q)$$

Table 6 enumerates seven examples (S4 – S9) with the following characteristics:

	F	*q*
S4	0	·9
S5	0	·5
S6	0	·1
S7	·9	·5
S9	·9	·1
S10	1·0	·615

For any value of *F* there is a continuum of strategies corresponding to a continuum of gene frequencies from 0 to 1. It is obviously not possible to try all possible two-way comparisons of these strategies to find the optimum mixture.

TABLE 6

A set of monomorphic and polymorphic strategies and their associated utilities in two states of nature. Strategies S1, S2, *and* S3 *are the homogeneous populations* AA, Aa *and* aa. *Strategies* S4 — S9 *are segregating populations with different gene frequencies and inbreeding coefficients.* S10 *is an optimal mixture of* AA *and* aa *homozygotes.*

	N1	N2
S1	8	0
S2	3	4
S3	2	7
S4	7·04	0·79
S5	4·00	3·75
S6	2·24	6·39
S7	7·36	0·71
S8	4·90	3·53
S9	2·56	6·31
S10	4·31	4·31

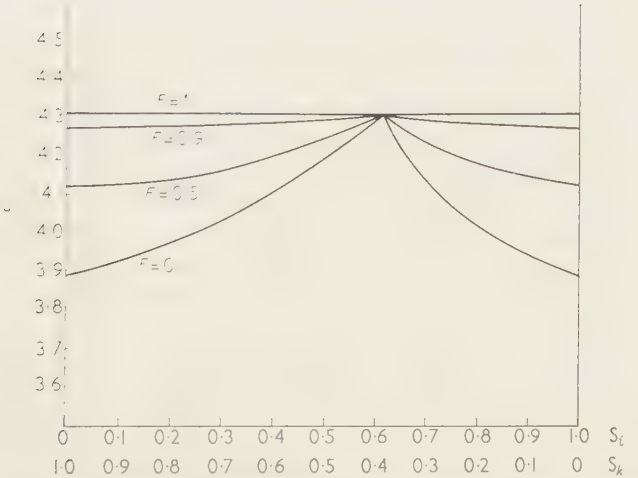

FIG. 4. The upper bounds of solutions to mixed strategies when the genotypes in Table 5 are assumed to be segregating in populations. Four bounds are shown corresponding to four levels of inbreeding, F = 0, ·5, ·9 and 1·0. The point at which they all meet is the maximin solution discussed in the text.

It can be shown, however, that the solution to the paired comparisons (the points of intersections of the utility lines) form a set bounded by a curve as shown in Fig. 4. Each curve in Fig. 4 is the upper bound of the set of solutions for a different value of inbreeding from $F = 0$ to $F = 1$. Each curve is in the form of two cusps with a singularity at $S_i = ·615$, $S_k = ·385$. The left hand limb of each curve happens to correspond to a mixture of S_i = homozygous *aa* and S_j segregating. The right hand limb

of each corresponds to S_i segregating and S_j = homozygous AA. The singularity for each curve then has the identical meaning: a mixture of ·615 *aa* populations and ·385 AA populations with *no* segregating populations. Since the singularity is the highest point with a utility of 4·31, the optimal strategy is one in which there is a mixture of homozygous AA and homozygous *aa* populations in proportions ·385 and ·615 respectively.

The line for complete inbreeding, $F = 1$, is particularly interesting. It is a horizontal straight line with the utility equal to 4·31 at every point. There is then no unique maximum solution for completely inbred populations. However, all the strategy mixtures on this line have following properties:

(1) There are no heterozygotes. This follows from the complete inbreeding, $F = 1$.
(2) The overall frequency of is *aa* ·615 and of AA is ·385 for every point on the line.

The second property is simply a reflection of the fact that completely inbred populations really consist of sub-populations each homozygous for AA or *aa*.

Thus the optimal strategy for this species consists in possessing only two genotypes *aa* and AA in the proportions ·615 to ·385. This can be accomplished by having all populations homogeneous within themselves, 40% being AA population and 60% *aa* populations. It can equally well be accomplished by inbreeding so that each population is a mixture of AA and *aa* homozygotes with an overall proportion of 40% AA to 60% *aa* in the species as a whole.

In biological terms, the situation of Table 6 favors *genetic isolation* between AA and *aa* genotypes. This means either speciation or complete selfing within one species. Which of these two alternatives is truly optimal depends upon the form of the game for other loci. Complete selfing leads to homozygosity of all the genes in the genome and this may not be optimal. Speciation on the other hand, creates an optimal condition for the new *species pair*, without forcing homozygosity at other loci.

The Choice of a Utility

I have deferred until last the question of an appropriate choice of a utility function because it is closely related to the question of the criterion of optimality. Population geneticists and ecologists are not in agreement about the parameter or parameters which measure the "success" of a population.

For many population geneticists the average fitness of a population is the appropriate choice. Dobzhansky (1951), Cain & Sheppard (1954) and Li

(1955) have discussed this point thoroughly and the result of that discussion may be quickly summed up. The average fitness of a population, $\overline{W}$, does tend to a local maximum under the forces of intra-populational selection in most simple cases. However, $\overline{W}$ is defined only within a population and does not, in itself, provide a standard of comparison between populations. All other things being equal, a higher $\overline{W}$ means greater "success", but, as always, other things are never equal. In some sense a population riddled with lethal genes is not as "successful" as one free of them since more of its zygotes are being wasted in the elimination of these genes. A large genetic load *is* a disadvantage to a population, but it is not the whole story. With some kinds of selection an increase in $\overline{W}$ within a population may actually lower the competitive ability of that population with respect to others of lower $\overline{W}$.

Among population ecologists the rate of increase, r, is popular as a measure of success. It is necessary to distinguish two meanings of r. First, there is the rate of increase of a population calculated as the birth rate minus the death rate at any instant of time. An r defined in this way is sensitive to population density and as MacArthur (1960) points out, r is identically zero for all populations which are stable in size, irrespective of that size. Birch (1960) uses the parameter r_0 (although not always), the rate of increase of a population *in its logarithmic growth phase*. The chief disadvantage to this measure is that it is *independent* of population size and measures only the returning power of a decimated population or the rate of increase in newly founded colonies. While these two phases in the life of a population are important, they cannot be the whole measure of success. A great advantage of r_0 is that it will be increased by natural selection since an individual who leaves more offspring leaves more genes. (This is not necessarily so if there is inter-family selection.) Any parameter under the control of a direct mechanism like natural selection has distinct advantages as a measure of population success.

Thoday (1953) has made the intriguing proposal that the *probability of survival* of a population or other evolutionary unit, is the best measure of fitness of that unit. The difficulty of such a parameter is that, as envisioned by Thoday it is virtually impossible to measure. Thoday considered the probability of survival to some distant, unspecified, time, but using the techniques of stochastic game theory, this concept may become more manageable. Consider a population playing a stochastic game extended in time. At each step or move, there is a probability P_{i0} that the population will contain no individuals at time $t + 1$, given that it contained i individuals at time t. These probabilities will be different for different strategies and states of nature and, in fact, correspond to the probability of playing

the terminal game Γ^0 referred to in a previous section. This extinction probability will be a function of both $\overline{W}$ and r and thus includes both genetic and ecological parameters. I propose, then that ($P_s = 1 - P_{i0}$), the probability of survival, is a parameter which it will be of advantage for a population to maximize over the whole stochastic game. Since states of nature are uncertain, however, this maximization must take the form of a maximin criterion. That is, the population's optimal strategy is to keep the *local probability* of survival as high as possible under the worst combination of states of nature. The application of the maximin principle to a stochastic game has not been discussed here, but methods are well-known. What is essential here is that if P_s is accepted as a measure of utility, then a maximin solution seems called for.

The Adoption of Strategies

The final element needed in a game theoretical approach to evolution is a *mechanism*. For intra-population genetics, the maximization of $\overline{W}$ is a good principle because, in fact, the processes of natural selection provide the mechanism for this maximization. Does a similar mechanism exist for "*maximinization*" of the probability of survival? The answer to this question depends entirely upon the importance of population and species extinction. The notion of natural selection is tautological in that it simply states that those individuals with the highest probability of survival are most fit. A similar tautology holds for population and species. Species which survive are by definition more fit than those which expire. Thus we expect that surviving groups will be those which have, *simply by chance*, acquired optimal strategies. This is completely analogous to the Darwinian notion of natural selection of chance variation on the individual level. But what makes natural selection go is the fact that every organism is mortal and the rate of individual extinction is very high. The rate of natural selection within a population is limited by the rate of individual extinction (barring exponentially increasing populations, which are rare). By the same token the effectiveness of inter-deme selection among strategies is bounded by the rate of extinction of demes.

The commonness of population extinction is subject to much discussion. Andrewartha & Birch (1954) believe the rate of extinction of demes to be high while others, notably Nicholson (1957) hold the opposite view. It is extremely difficult to get data on extinction rates of species or higher taxonomic categories because of the confusion of phyletic extinction with taxonomic extinction, the changing of a taxonomic name due to phyletic evolution. Really intensive examination of certain aspects of the fossil record on the lines of Simpson's elucidation of the Equids would provide the necessary information. If the horse were typical, which is unlikely,

the rate of extinction of distinct phyla would be quite high. Living Equid phyla represent less than 1% of all known phyla since the Eocene.

On the side of experimentation much can be done. Although the time spans involved in selection among populations is obviously much greater than among individuals, it does not follow that the course of intra-population selection cannot be followed in the laboratory. For example large numbers of experimental bottle or vial populations can be kept simultaneously with organisms like *Tribolium* or *Drosophila*. If different populations were allowed different pure strategies (homozygosity for different alleles, polymorphism, different amounts of recombination, etc.) the extinction of these populations could be followed in fluctuating environments. Mimicking of natural fluctuation in temperature, say, would be relatively easy.

In short, experimentation and observation will reveal the kinds of strategies that promote the longevity of populations and species and thus define *biologically* the meaning of an optimal strategy.

REFERENCES

Andrewartha, H. G. & Birch, L. C. (1954). "The Distribution and Abundance of Animals". University of Chicago Press.

Birch, L. C. (1960). *Amer. Nat.* **94,** 5.

Carson, H. L. (1955). *Cold Spr. Harb. Symp. quant. Biol.* **20,** 276.

Carson, H. L. (1959). *Cold Spr. Harb. Symp. quant. Biol.* **24,** 87.

Cain, A. J. & Sheppard, P. M. (1954). *Amer. Nat.* **88,** 321.

Dobzhansky, Th. (1951). "Genetics and the Origin of Species". 3rd Edition. Columbia University Press.

Dobzhansky, Th. (1955). *Cold Spr. Harb. Symp. quant. Biol.* **20,** 1.

Dobzhansky, Th. & Levene, H. (1955). *Genetics,* **40,** 797.

Fisher, R. A. (1930). "The Genetical Theory of Natural Selection." Clarendon Press, Oxford.

Lewontin, R. C. (1957). *Cold Spr. Harb. Symp. quant. Biol.* **22,** 395.

Li, C. C. (1955). *Amer. Nat.* **89,** 281.

Luce, R. D. & Raiffa, H. (1957). "Games and Decisions". John Wiley & Sons, New York.

MacArthur, R. H. (1960). *Amer. Nat.* **94,** 313.

Mayr, E. (1959). *Cold Spr. Harb. Symp. quant. Biol.* **24,** 1.

Nicholson, A. J. (1957). *Cold Spr. Harb. Symp. quant. Biol.* **22,** 153.

Thoday, J. M. (1953). *Symp. Soc. exp. Biol.* **7,** 96.

von Neumann, J. & Morgenstern, O. (1944). "Theory of Games and Economic Behavior". 1st Edition. Princeton University Press.

Wallace, B. (1959). *Cold Spr. Harb. Symp. quant. Biol.* **24,** 193.

Wright, S. (1931). *Genetics,* **16,** 97.

J. Theoret. Biol. (1961) **1**, 404–406

Wave Mechanics and Muscular Contraction

WILLIAM ARNOLD

Biology Division, Oak Ridge National Laboratory,†
Oak Ridge, Tennessee, U.S.A.

(*Received 27 March* 1961)

The exact mechanism whereby the muscle transforms chemical into mechanical energy is not known, although "theories of muscular contraction are available on the market by the dozen" (Szent-Györgyi, 1957). It is not our purpose here to add a new theory to this list but to answer the question: *Are quantum mechanical forces, by themselves, large enough to account for the tension developed by a muscle?*

We answer our question by calculating how large a contractile unit would have to be for the quantum mechanical forces to give the observed tension. We ignore all other forces.

Let us assume that a muscle contains contractile units and that:

1. Each unit is a rectangular parallelepiped with the dimensions Z, Z, X in centimeters.
2. The unit can change its shape, but the volume is constant:

$$Z \times Z \times X = X_0^3 \tag{1}$$

3. An electron or hole given to the unit is free to move about inside the unit. It is to be this act that starts the contraction.
4. In the relaxed state of the unit $X > Z$.

The lowest energy level for an electron or hole in a box (Kauzmann, 1957) is given by

$$E = \frac{h^2}{8m}\left(\frac{1}{Z^2} + \frac{1}{Y^2} + \frac{1}{X^2}\right) \tag{2}$$

where X, Y, and Z are the dimensions, h is Planck's constant, and m is the mass of the electron.

For our unit, where $Z = Y$ and the volume is constant, equation (2) can be written as

$$E = \frac{h^2}{8m}\left(\frac{2X}{X_0^3} + \frac{1}{X^2}\right) \tag{3}$$

† Operated by Union Carbide Corporation for the U.S. Atomic Energy Commission.

Since $\frac{dE}{dX}$ is positive, there is a force F tending to make the unit into a cube, that is, to contract it in the direction of X,

$$F = \frac{dE}{dX} = \frac{h^2}{8m}\left(\frac{2}{X_0^3} - \frac{2}{X^3}\right) \tag{4}$$

In Fig. 1 this force is plotted as a function of X/X_0. For comparison, the force obtained from single muscle fibers is plotted as a function of length;

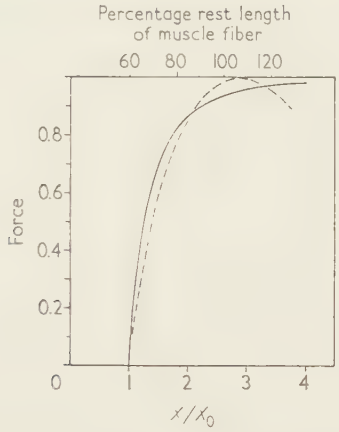

FIG. 1. Force-length diagram. The solid line gives the force as a function of length for the contractile units. The dashed line gives the force as a function of length for a single muscle fiber. For each curve the maximum force has been set equal to one.

the curve was taken from the length-tension diagram (Huxley, 1960a) by assuming the force to be the tension divided by the length. The two curves were made to coincide at $F = 0$. The scale was chosen to make them look somewhat alike.

From the figure it would seem that a value of $X/X_0 = 2$ to 3 for the relaxed unit would be reasonable. We shall use the value of 3 in our calculations. As the unit starts to contract the tension is given approximately by

$$T = \frac{F}{Z^2} = \frac{h^2 6}{8mX_0^5}$$

since

$$Z^2 = \frac{X_0^3}{3X_0} \tag{5}$$

We take the maximum tension that a muscle can exert to be 2 kg per sq cm (Huxley, 1960b). We assume that the contractile units make up half

of the cross section of the muscle, so that the tension in the unit will be

$$T = \frac{h^2 6}{8mX_0^5} = 4 \times 10^6 \text{ dynes per cm}^2 \tag{6}$$

Then

$$X_0 = 3{\cdot}9 \times 10^{-7} \text{ cm} = 39 \text{ Å} \tag{7}$$

Knowing X_0 and taking the density of the unit to be ρ, we calculate that the molecular weight of the contractile unit is $36{,}000 \times \rho$. This value agrees with the estimate made by Szent-Györgyi (1953) that a unit of actomyosin of molecular weight 35,000 to 70,000 converts the energy of one ATP molecule into work.

During contraction the energy level of the electron, or hole, drops by an amount ΔE approximately

$$\Delta E = \frac{h^2 3}{8mX_0^2} \tag{8}$$

or

$$\begin{aligned} \Delta E &= {\cdot}12 \times 10^{-12} \text{ ergs} \\ &= {\cdot}074 \text{ electron volts} \\ &= 1700 \text{ calories per mole} \end{aligned}$$

The energy ΔE could be easily furnished by one ATP molecule.

Thus, the quantum mechanical forces involved in confining an electron, or hole, to a volume corresponding to a molecular weight of 36,000 are able to account for the tension developed by muscle.

REFERENCES

Huxley, H. E. (1960a). "The Cell", Vol. 4, Pt. 1, 413. (Jean Brachet and Alfred E. Mirsky, Eds.) Academic Press, New York.

Huxley, H. E. (1960b). "The Cell", Vol. 4, Pt. 1, 448. (Jean Brachet and Alfred E. Mirsky, Eds.). Academic Press, New York.

Kauzmann, W. (1957). "Quantum Chemistry", 187. Academic Press, New York.

Szent-Györgyi, A. (1953). "Chemical Physiology of Contraction in Body and Heart Muscle", 36. Academic Press, New York.

Szent-Györgyi, A. (1957). "Bioenergetics", 59. Academic Press, New York.

J. Theoret. Biol. (1961) 1, 407–409

Letter to the Editor

The Effect of Concentration and Rate of Intravenous Injection of Thiopentone

Several years ago one of us (L.N.) made a short study of the ketosis occurring during anaesthesia, and found that an I.V. injection of thiopentone, given for routine surgical procedures, quite often produced a significant rise in ketone bodies within a minute to a minute and a half of the injection. On attempting to repeat and confirm these observations in recent months it has not proved possible to show that such a ketosis occurs at all regularly, and even when it does occur the changes noted are small. This difference in the effect of I.V. thiopentone was attributed to the change in anaesthetic technique which had occurred during the intervening years, a change due mainly to the danger of arterial spasm and the resultant possibility of gangrene if an injection of a 5% solution of thiopentone were accidentally made into an aberrant artery instead of into a vein. The concentration of thiopentone now routinely used is 2·5% and the dose of this material is injected more slowly.

We have found it of interest, as a result of Professor Paton's paper (1960) to attempt a quantitative expression of the results of reducing the concentration of the injected thiopentone and at the same time lengthening the period of the injection of the drug. This attempt is a first approximation only.

If the dose administered to a man weighing 70 kg were 0·9 gm thiopentone, i.e. 18 ml of 5%, and this were initially distributed evenly throughout the extracellular volume of 18 litres, the concentration produced would be 0·005%, and this would be the concentration level at which it would be assumed that the action of the drug was being observed. If, however, "slug" formation occurs, as Professor Paton has suggested, the rapid I.V. injection of the thiopentone would result in a "concentration-front" reaching the tissue cells at the end of the first circulation. A necessary assumption is that, unless the drug under consideration has a special affinity for lung tissue, the "slug" having passed through the right heart and pulmonary circulation, re-forms without very great dispersion to pass through the left heart to reach the systemic capillary bed. In this first circulation the "slug", originally an axial stream of 18 ml (of 5% as injected) has been diluted by a volume of about 4 stroke volumes, i.e. 18 ml in 288 ml or

1/16th of the starting concentration. Further dilution in the extra-cellular space would bring the "concentration-front" reaching the tissue cell surface to about 0·1%. This is 20 times greater than the concentration assumed by a calculation based on the immediate and even distribution of the drug throughout the extra-cellular space.

However, not only will the total amount of the drug entering an accepting cell be increased in proportion to the 20-fold concentration existing during the first circulation, but the rate of entry is also affected. It can be shown that, to a first approximation, where a cell has a membrane permeable to a given solute, the rate at which the cell may take up the solute will be proportional to the logarithm of the concentration gradient between cell and environment. In the example given, the ratio of the rate of entry of thiopentone into the cell (assuming immediate even distribution) to the rate of entry (assuming the concentration-front of a "slug") is log 5/log 100, or a factor of approximately $\frac{1}{3}$. Thus, with "slug" formation, both the total amount of the drug entering a permeable cell wall and the rate of entry are very greatly increased over the values assumed by considering only the equilibrium distribution.

A further point is that the portion of the solute to be re-excreted unmetabolized by the cell to the environment (in this case thiopentone temporarily dissolved in the adipose organ) will be released at a rate again governed, at least in part, by the logarithm of the gradient, now from cell to environment. Where previously a high environmental concentration was presented to a cell concentration originally zero (for the solute), in reverse the high cell concentration presents to an environment approaching the so-called equilibrium value. Hence the rate of exit of a drug from a cell will be slower than its entry by a factor related to the rate of I.V. administration.

The relevance of these considerations to the use of thiopentone for anaesthesia is apparent from the dual pharmacological action ascribed to barbiturates. In light barbiturate anaesthesia breathing is regulated by CO_2 tension on the respiratory centre, i.e. CO_2 can be used as a stimulator of respiration. In deep barbiturate anaesthesia, on the other hand, regulation by CO_2 is abolished and the stimulus is hypoxia, CO_2 in these circumstances intensifying the barbiturate depression of the centre. It would seem likely that the earlier technique of thiopentone administration produced mainly the deep form of barbiturate anaesthesia due to "slug" formation. In previous years surgical procedures were carried out under thiopentone anaesthesia alone whereas it is now known that sleep-doses of thiopentone actually lower the pain threshold.

From this example we conclude that the precise pharmacology of I.V. administered drugs may require careful re-scrutiny in individual cases to

determine which concentration level and rate of cell entry was, in actual fact, being studied.

St. George's Hospital
Long Leys Road
Lincoln, England

L. NAFTALIN
A. STEPHENS

(*Received* 28 *March* 1961)

We wish to thank the M.R.C. for a grant for technical assistance for work on post-operative ketosis; this letter reports one aspect of the work.

REFERENCE

PATON, W. D. M. (1960). *Proc. roy. Soc. Med.* **53**, 815.

Instructions to Authors

1. The *Journal of Theoretical Biology* is intended to be a central publishing medium for theoretical work in all fields of biology.

Appropriate subjects for papers will include:

(*a*) generalized theories;
(*b*) theories of specific processes or phenomena, e.g. crossing over, learning, active transport, photosynthesis, carcinogenesis;
(*c*) theoretical discussion of specific projects, e.g. selective mutation, drug design, electronic models;
(*d*) theoretical discussion of methods, e.g. cytochemical methods and instrumentation, amplification techniques;

Papers concerned mainly with experimental work will not normally be accepted, but a paper will not be rejected simply because it contains some new experimental material.

2. Reviews will be accepted only if they include a substantial new theoretical discussion. Letters to the editor, intended to direct attention to a particular theoretical point, will be accepted.

3. *Submission of Manuscripts.* Manuscripts should be sent to Professor J. F. Danielli, Zoology Department, King's College, London, or to any member of the Editorial Board. The submission of manuscript will be taken to imply that the material is original.

Papers may be submitted in English, French or German. It is emphasized however that papers in English will reach a much larger group of readers than those in other languages. A paper submitted in French or German should have an abstract in English.

Publication will be accelerated if two copies of a manuscript are submitted.

4. *Manuscripts.* If the Journal is to serve its intended purpose, it is essential that the main body of its contents should be intelligible to all biologists. To this end, wherever possible mathematical development should be confined to appendices, and specialized terms should be clearly defined in the introduction to a paper.

Summary. Every paper should be preceded by a summary: this should not normally exceed 300 words, and should be intelligible to the general reader without reference to the main text. Abbreviations should not be used.

Lay-out. Manuscripts should be typewritten in double spacing throughout (including summary, footnotes, tables and legends) on one side of quarto paper, with wide margins. Footnotes, tables and legends for illustrations should be typed separately at the end of the manuscript.

Running Headline. A short running title of not more than 40 characters (including spaces), suitable for page headings, should be given if the full title is longer than this.

5. *Conventions.* The conventions used in current issues of the Journal for headings, references, etc., should be followed in preparing manuscripts. References should be listed alphabetically at the end of the paper, and abbreviated according to the *World List of Scientific Periodicals.* Citations in the text should read thus: Smith, Brown & Robinson (1957), or (Smith, Brown & Robinson, 1957). When a citation has more than two authors, the names of all of them should be given at the first citation (except for more than five authors), and on subsequent citations thus:

(Smith *et al.*, 1957). The conventions (Brown, 1957a), (Brown, 1957b) should be used when more than one paper by the same author(s) has appeared in one year. Volume number should be indicated in the typescript thus: 53.

Footnotes should be avoided whenever possible. Customary abbreviations will be used; authors may refer to *Chemical Abstracts*, to the *Biochemical Journal*, **66**, 8 (1957), and to *Notes on the preparation of papers to be communicated to the Royal Society* (1951). Certain non-standard abbreviations of chemical substances are convenient, and those listed in the *Journal of Biological Chemistry* may be used without further definition (e.g. DNA, ATP). Any other abbreviations of this kind should be used sparingly, and must be defined in a footnote at the point in each manuscript where the first such abbreviation occurs.

6. *Illustrations*. It is often convenient to make a drawing two or three times larger than it will appear in the Journal, but very large drawings should be avoided and it will be helpful if drawings and illustrations are submitted on quarto sheets. Drawings should be made in Indian ink on tracing linen, Bristol board, or faintly blue-lined graph paper; the originals should always be submitted with the manuscript, since photo-prints of drawings cannot be reproduced satisfactorily.

Line drawings will be referred to as Fig. 1, Fig. 2, etc., and photographs as Plate I, Plate II, etc. The name(s) of the author(s) and the number should be written on all illustrations, together with the indication "Top" if necessary. Besides the original illustrations *one* duplicate set must accompany the second copy of the manuscript; these may be photo-prints, and need not be of the same quality as the originals but must be clear and legible. If oversize original drawings are submitted (bigger than 8 by 10 inches) they must be accompanied by *two* sets of small photo-prints.

7. *Reprints*. Fifty reprints of each paper will be supplied free of charge; additional copies may be purchased. Order forms will be sent with proofs.

In preparation

HORMONES IN BLOOD

Edited by

C. H. GRAY, *King's College Hospital Medical School, London*
and A. L. BACHARACH, *Formerly of Glaxo Laboratories Ltd.*

July 1961, 654 pages, approx. 120s.

CONTENTS:

ACADEMIC PRESS, London and New York

Volume 1, Number 1, Autumn 1961

EXPERIMENTAL EYE RESEARCH

A new international journal
devoted to scientific research on the eye

Volume 1, 1961/1962 (issued quarterly), 115s.

Editors: HUGH DAVSON, *University College, London*
and ENDRE A. BALAZS, *Retina Foundation, Boston, Massachusetts*

Experimental Eye Research will be available for the publication of the results of original research on the anatomy, physiology, biochemistry and biophysics of the eye. Contributions on such subjects as the production and circulation of ocular fluids, the fine structure, biochemistry, biophysics and metabolism of eye tissues, the photochemical processes related to visual function, and the neuromuscular mechanisms in the control of eye movements will be preferred.

Papers will be welcomed from all parts of the world, the sole criterion for publication being that they be written in English and that their scientific content meet the high standards imposed by editorial policy.

Experimental Eye Research will be published quarterly and will contain approximately 400 pages per volume.

Manuscripts should be sent to either Editor.

Subscription orders should be sent to the publishers.

ACADEMIC PRESS, London and New York

Journal of Molecular Biology

Published at 128s. 6d. ($18.00) per annual volume of 6 issues

Volume III Number 3, June 1961

CONTENTS

ACADEMIC PRESS INC.
LONDON NEW YORK

Contents

Printed in Great Britain by The Whitefriars Press Ltd., London and Tonbridge